The Open University

M248

Analysing data

Computer Book B

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2017.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Hobbs the Printers Limited, Brunel Road, Totton, Hampshire SO40 3WX.

ISBN 978 1 4730 2263 8

5.1

Contents

Introduction

Using this book

As you study each unit in Book B, you will be directed to work through particular chapters in this book as part of your work on that unit. Each unit contains instructions as to when you should first refer to particular material in this computer book; you are advised not to work on the activities here until you have reached the appropriate points in the units.

As with Computer Book A, the activities vary in nature and length. Some contain instructions on how to use the module software to perform particular tasks; some contain instructions on how to use the module animations to investigate statistical ideas or to help your understanding of concepts. Yet others provide practice at using the software to explore or analyse data; you will find solutions to these activities at the end of this computer book. You should try to work through all the activities as you read the chapters.

A few supplementary exercises on the whole of this computer book are provided after Chapter 11. You may use these for extra practice or for revision (or not at all), as you wish.

1 Normal distributions

This chapter is associated with Section 2 of Unit 6.

In Computer Book A, you used Minitab to find probabilities and quantiles for a number of distributions including the binomial, Poisson and exponential distributions. You will recall that this is done using **Probability Distributions** from the **Calc** menu. In this chapter, you will use this Minitab facility to find probabilities and quantiles for normal distributions and briefly investigate some properties of normal distributions.

Activity 1 *Chest measurements of Scottish soldiers*

In Unit 6, a normal distribution with mean 40 and standard deviation 2 was used to model the variation in the chest measurements (in inches) of Scottish soldiers in the nineteenth century. In this activity, you will use Minitab to calculate several probabilities for this distribution.

These probabilities were calculated by hand in Examples 4 and 5 of Unit 6.

Start Minitab now if it is not already open.

- Select **Calc > Probability Distributions > Normal. . . .**

The **Normal Distribution** dialogue box will open. This dialogue box is similar to those that you have already met. The only difference is in the information that you need to enter in order to specify which member of the family of normal distributions is required. For a normal distribution, you need to enter values for the mean and standard deviation.

Note that Minitab requires the standard deviation of the normal distribution, not the variance.

- Enter 40 in the **Mean** field and 2 in the **Standard deviation** field.

See, for example, Activities 28 and 45 in Computer Book A.

Recall from Computer Book A, that values of the c.d.f. are found using **Cumulative probability**, and quantiles are found using **Inverse cumulative probability**. You will need to use **Cumulative probability** to answer parts (a) and (b) below, and **Inverse cumulative probability** to answer parts (c) and (d).

(a) Use Minitab to calculate the proportion of Scottish soldiers who, according to this model, had chests measuring between 37 and 42 inches inclusive.

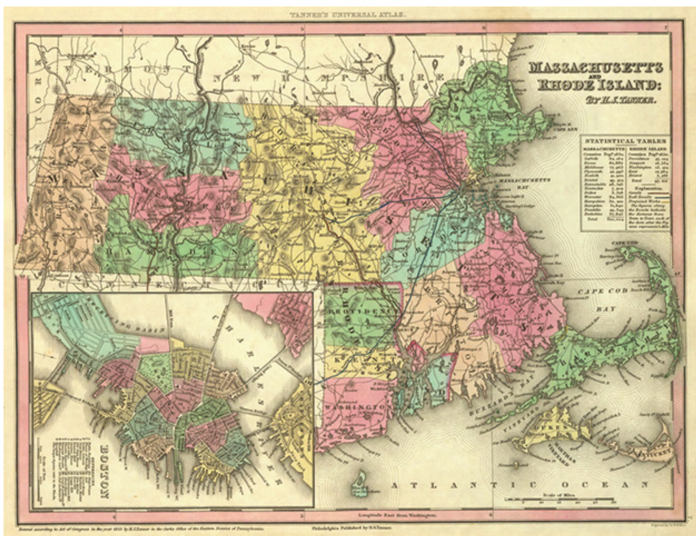
(b) According to this model, what proportion of Scottish soldiers had chest measurements greater than 43 inches?

(c) According to this model, below what value were the chest measurements of only 2.5% of Scottish soldiers?

(d) What chest measurement was exceeded by only 5% of Scottish soldiers according to this model?

Activity 2 *Errors in angular measurements*

In Activities 6 and 7 of Unit 6, a normal distribution with mean 0 and variance 2.75 was used to model angular measurement errors made during the mapping of the state of Massachusetts in the USA in the nineteenth century. In this activity you should use Minitab to find the probabilities and quantiles discussed in those activities as listed below. (Round the value for the standard deviation that you use to three decimal places.)



Map of Massachusetts, 1837

(a) Find the probability that an error is positive and greater than 0.5 minutes of arc.

Calc > Probability Distributions > Normal...

(b) What is the probability that an error is positive but less than two minutes of arc?

(c) Obtain the probability that the size of an error, which may be positive or negative, is less than 1 minute of arc.

(d) What is the probability that the size of an error, positive or negative, is greater than 1.5 minutes of arc?

(e) Suppose that 10% of errors are positive and greater than x. Find the value of x.

(f) Suppose that 95% of errors lie between $-b$ and b (where b is positive). Find the value of b.

(g) Suppose that 99% of errors lie between $-c$ and c (where c is positive). Find the value of c.

Activities 3 and 4 will give you some further practice at deciding which probabilities or quantiles are required to solve a problem, and at using Minitab to find their values. Draw rough sketches to help you decide which values are required.

Activity 3 *Heights of elderly women*

Suppose that the variation in the heights (in centimetres) of elderly women may be adequately modelled by a normal distribution with mean 160 and standard deviation 6.

(a) According to the model, what proportion of elderly women are taller than 166 cm?

(b) Find the probability that the height of a randomly selected woman will be between 145 cm and 157 cm.

(c) Below what value are the heights of 85% of elderly women?

(d) Find the interquartile range of the heights of the population of elderly women.

In May 2015, Harriette Thompson became the oldest woman to complete a 26.2-mile marathon, at 92 years of age (*Washington Post*, 31 May 2015)

Activity 4 *Nicotine levels of smokers*

Suppose that the variation in the blood plasma nicotine levels (in nanograms per millilitre) of smokers may be modelled by a normal distribution with mean 315 and standard deviation 131.

(a) According to the model, what proportion of smokers have nicotine levels lower than 300?

(b) What proportion of smokers have nicotine levels between 150 and 400?

(c) What nicotine level is such that only 20% of smokers have a higher level?

In Unit 6 it was stated that, for a normal distribution, observations more than three standard deviations from the mean are rather unlikely. In the next activity you are asked to investigate this using Minitab.

Activity 5 *Observations more than three standard deviations from the mean*

(a) In Activity 1, a normal distribution with mean 40 and standard deviation 2 was used to model the variation in the chest measurements of nineteenth-century Scottish soldiers. According to the model, what proportion of Scottish soldiers had chest measurements more than three standard deviations from the mean, that is, below 34 or above 46?

(b) In Activity 2, a normal distribution with mean 0 and variance 2.75 was used to model errors in angular measurements. According to the model, what proportion of errors are more than three standard deviations from the mean?

(c) In Activity 3, a normal distribution with mean 160 and standard deviation 6 was used to model the variation in the heights of elderly women. According to the model, what proportion of elderly women are either more than three standard deviations shorter or more than three standard deviations taller than the mean height?

(d) What do you notice about your results in parts (a) to (c)? Conjecture a general result for normal distributions and 'test' your conjecture for a pair of values of the mean μ and standard deviation σ of your choice.

In Activity 5 you found that, whatever the values of the mean and standard deviation of a normal distribution, the proportion of observations more than three standard deviations from the mean is the same and is approximately equal to 0.0027. So, the proportion of observations within three standard deviations of the mean is 0.9973. What proportion of observations are within one standard deviation of the mean, or within two standard deviations of the mean? Are the proportions the same for all normal models? You will investigate these questions in Activity 6.

Activity 6 *Observations within k standard deviations of the mean*

(a) For each of three pairs of values of the mean μ and the standard deviation σ of your own choice, find the following proportions.

- The proportion of observations within one standard deviation of the mean.
- The proportion of observations within two standard deviations of the mean.

Summarise your results.

(b) For each of the normal models you used in part (a), find the proportion of observations within k standard deviations of the mean, where k is a positive value of your own choice.

(c) Conjecture a general result for normal distributions on the basis of what you have observed in parts (a) and (b).

2 Normal probability plots

This chapter is associated with Section 5 of Unit 6.

In Section 5 of Unit 6, you have seen how to construct a probability plot to check whether a normal distribution is a plausible model for the variation in a dataset. In this chapter, you will use Minitab to obtain normal probability plots for several datasets.

There are different ways of deciding which quantiles to use when constructing a probability plot; Minitab offers four methods. The method that is described in M248 is not the one Minitab produces as a default, so you will need to change the options.

Start Minitab now if it is not already open.

Activity 7 *Changing the option settings*

- Use **Tools** > **Options...** to open the **Options: General** dialogue box.
- Click on the **+** to the left of **Individual Graphs** (on the left-hand side) to display a list of graphs.
- Click on **Probability Plots** to view the options available for probability plots.
- Under **Y-Scale Type**, select **Score**.
- Under **Graph Orientation**, make sure that **Show raw data on horizontal scale** is selected.
- Next, you must specify the method to be used to obtain the plot points. The method described in Unit 6 is obtained by selecting **Mean Rank (Herd-Johnson)**. Click on this to select it.
- Click on **OK** to close the dialogue box.

The settings you have selected will be used for all probability plots that you produce unless you change them. Some of the settings can be changed when you produce an individual probability plot. However, the method for calculating the plot points can be changed only using **Tools** > **Options...**.

You are now ready to use Minitab to produce a normal probability plot.

Activity 8 *Silver content of coins*

Data on the silver content of twelfth-century Byzantine coins minted during different periods of the reign of Manuel I Comnenus are in the worksheet **coins.mtw**. Is it reasonable to assume that, for each coinage, the data are observations from a normal distribution? In this activity, you will investigate whether a normal distribution is a plausible model for the silver content of coins from the first coinage.

Map of Byzantine Empire around 1180 at the end of the reign of Manuel I Comnenus

Open the worksheet **coins.mtw** now. A normal probability plot for the data on silver content of coins from the first coinage is obtained as follows.

- Select **Graph > Probability Plot...** . The **Probability Plots** dialogue box will open.
- Select **Single** (the default) and click on **OK** to open the **Probability Plot: Single** dialogue box.
- The data for the first coinage are in column `Coin1`, so enter `Coin1` in the **Graph variables** field.
- Click on the **Distribution...** button to open the **Probability Plot: Distribution** dialogue box.
- Check that `Normal` is displayed in the **Distribution** field of the **Distribution** panel (this is the default option). If `Normal` is not displayed, then select it from the **Distribution** drop-down list.
- Click on the **Data Display** tab to view the **Data Display** panel.
- When displaying a probability plot, by default Minitab also includes 'confidence intervals' on the plot. This feature is not required here. Click on **Show confidence interval** to switch off this feature.
- Click on **OK** to close the **Probability Plot: Distribution** dialogue box.
- Click on **OK** in the **Probability Plot: Single** dialogue box and a normal probability plot will be displayed: this is Minitab's version of Figure 29 of Unit 6.

The normal probability plot displayed by Minitab includes some summary statistics in a box to the right of the plot. Figure 1 shows the normal probability plot after this box and the title have been removed.

The summary box and title will be removed from all normal probability plots in M248 from now on without comment.

The points lie roughly along a straight line, so a normal distribution is a plausible model for the silver content of coins from the first coinage.

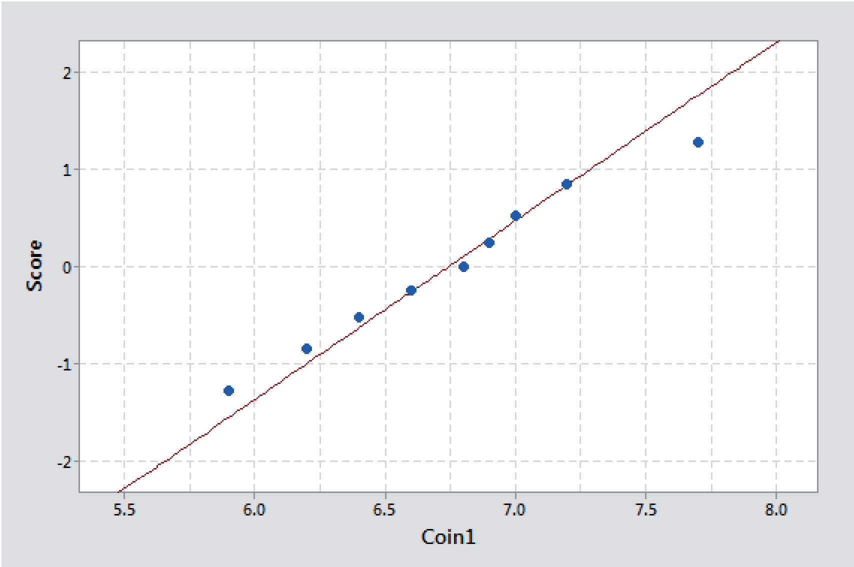


Figure 1 A normal probability plot for the first coinage

Activity 9 *Nicotine levels of smokers*

Data on the blood plasma nicotine levels of 55 smokers are contained in the worksheet **plasma.mtw**. Obtain a normal probability plot for these data. Is a normal distribution a plausible model for the variation in blood plasma nicotine levels?

Data provided by D.J. Hand, Imperial College London.

Graph > Probability Plot...

Activity 10 *Shoshoni rectangles*

The ancient Greeks called a rectangle 'golden' if the ratio of its width to its length was $\frac{1}{2}(\sqrt{5}-1) \simeq 0.618$. The Shoshoni, a tribe of native North Americans, used beaded rectangles to decorate their leather goods. The data in the worksheet **shoshoni.mtw** are the width-to-length ratios for twenty of their rectangles. They were analysed as part of a study in experimental aesthetics.

Is a normal distribution a plausible model for the variation in the ratios?

This Shoshoni bag, dated 1927, is kept in the National Museum of American History, Smithsonian Institution

DuBois, C. (1960) *Lowie's Selected Papers in Anthropology*, University of California Press, pp. 137–42.

3 Examining and comparing estimators

This chapter is associated with Subsection 1.3 of Unit 7.

In this chapter, you will use two of the M248 animations to examine and compare estimators for two sampling models: a Poisson distribution in Activity 11 and a normal distribution in Activity 12.

Activity 11 *Comparing estimators of the Poisson parameter*

The mean and variance of a Poisson distribution with parameter λ are both equal to λ. Hence, given a sample from a Poisson distribution, either the sample mean $\overline{x}$, or the sample variance s^2, might be used as an estimate of λ. That is, random variables $\overline{X}$ (the sample mean) and S^2 (the sample variance) may both be used as estimators of λ. In this activity, you will use the animation **Estimators of the Poisson parameter** to compare these two estimators.

- Open the **Estimators of the Poisson parameter** animation.

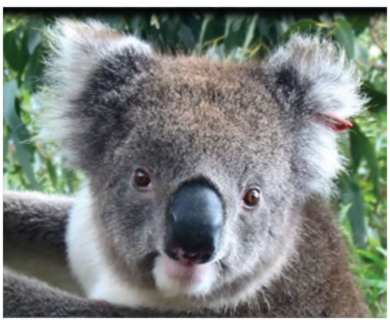
The animation simulates taking a large number (N) of samples, each of size n, from a Poisson distribution with parameter λ. The p.m.f. of the Poisson(λ) distribution is shown on the left-hand side of the animation. The controls for the simulation are below this. There are fields for the parameter λ, the sample size n and the number of samples to be drawn N. The default settings of λ, n, and N are 1, 10 and 1000, respectively.

You should use the default settings for parts (a), (b) and (c) of this activity.

(a) For each individual sample (of size n), the animation calculates the observed sample mean $\overline{x}$, and the observed sample variance s^2. We therefore have a sample of N values $\overline{x}_1, \overline{x}_2, \dots, \overline{x}_N$ of $\overline{X}$, and N values $s_1^2, s_2^2, \dots, s_N^2$ of S^2. When the simulation is run, unit-area histograms of these samples will be plotted on the right-hand side of the animation (initially these two graphs are empty): the histogram of the sample means (i.e. $\overline{x}_1, \overline{x}_2, \dots, \overline{x}_N$) will appear in the top right-hand graph, while the histogram of the sample variances (i.e. $s_1^2, s_2^2, \dots, s_N^2$) will appear in the bottom right-hand graph.

- Click on the **Take sample** button to run the simulation.

Look at the shapes of the two histograms produced by the simulation. Is either histogram symmetric? Does either of them show clear skewness?



When we typed 'Poisson' into Google Images we got lots of pictures of fish, and eventually the picture of Siméon-Denis Poisson in Unit 3 and this picture of a koala. The latter arises because the image has been improved by use of a sophisticated algorithm with other mathematical work of S.-D. Poisson at its heart.

(b) If $\overline{X}$ is an unbiased estimator of λ, then the mean of the sample means (i.e. the mean of $\overline{x}_1, \overline{x}_2, \dots, \overline{x}_N$) should be close to the value of λ. Similarly, if S^2 is an unbiased estimator of λ, the mean of the sample variances (i.e. the mean of $s_1^2, s_2^2, \dots, s_N^2$) should also be close to the value of λ. So, in order to assess whether $\overline{X}$ and S^2 are unbiased estimators of λ, the animation calculates the mean of the sample means, and the mean of the sample variances. These means are given underneath each respective histogram.

A large difference between the mean of the sample means and λ would indicate that $\overline{X}$ is a biased estimator of λ, whereas a small difference may be attributable to random variation, in which case the estimator may be unbiased. Similarly a large difference between the mean of the sample variances and λ would indicate that S^2 is a biased estimator of λ, while if there is a small difference, S^2 may be unbiased.

- Note down the mean of the sample means, and the mean of the sample variances, in the first empty column (labelled '1') of Table 1 below.
- Complete Table 1 by clicking on the **Take sample** button four more times, noting down the mean of the sample means, and the mean of the sample variances, each time.

Table 1 Observed means for 1000 samples, each of size 10, from Poisson(1)

Simulation number	1	2	3	4	5
Mean of the sample means					
Mean of the sample variances					

What value does the mean of the sample means seem to centre at? What value does the mean of the sample variances seem to centre at? Do you think one or both of $\overline{X}$ and S^2 could be an unbiased estimator of λ?

(c) If two estimators are unbiased, then the one with the smaller variance will tend to give more precise estimates than the other estimator. In order to assess the variances of the two estimators $\overline{X}$ and S^2, the animation calculates the variance of the sample means (i.e. the variance of $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N$) and the variance of the sample variances (i.e. the variance of $s_1^2, s_2^2, \ldots, s_N^2$), and then calculates the ratio of the variances to compare them, i.e. the animation calculates:

$$\text{Ratio of variances} = \frac{\text{Variance of the sample variances}}{\text{Variance of the sample means}}.$$

If this ratio is 1, this would suggest that the variances of the two estimators are the same. However, if this ratio is greater than 1, this would suggest that the variance of S^2 is larger than the variance of $\overline{X}$, while if the ratio is less than 1, this would suggest that the variance of S^2 is smaller than the variance of $\overline{X}$. The variance of the sample means, and the variance of the sample variances, are both given underneath each respective histogram (on the right-hand side), while the ratio of variances is given underneath the plot of the Poisson p.m.f. on the left-hand side.

- Complete Table 2 by clicking on the **Take sample** button five times, noting down the value of the ratio of the variances each time.

Table 2 Observed ratio of variances for 1000 samples, each of size 10, from Poisson(1)

Simulation number	1	2	3	4	5
Ratio of variances					

Which estimator seems to have the larger variance? Would you rather use $\overline{X}$ or S^2 as an estimator of λ when n is 10 and λ is 1?

(d) In part (c), you may have noticed that the values of the ratio of variances varied quite a lot. You can reduce the variability of this ratio by substantially increasing the number of samples that are generated in each simulation. So set the **Number of samples, *N*** field to a large value. (For most computers, 10 000 is a reasonable choice, but use a smaller value if the animation runs too slowly for you.)

Run the simulation (by clicking on the **Take sample** button) for sample sizes $n = 5$, 10, 50 and parameter values $\lambda = 1$, 3, 9, 27. As you run the simulations, do the following.

- Observe the shapes of the histograms for the sample means and for the sample variances. What do you notice about the way the shape of the histogram for the sample means changes as n is varied? Had you expected it to change as it did? What about the shape of the histogram for the sample variances?
- For each simulation, compare the mean of the sample means with the mean of the sample variances.
- For each simulation, enter the observed value of the ratio of variances in Table 3.

Table 3 Observed ratio of variances

n	$\lambda = 1$	$\lambda = 3$	$\lambda = 9$	$\lambda = 27$
5				
10				
50				

Does there seem to be a pattern in the way the ratio of variances varies with n and λ?

On the basis of all the simulations you have run, do you think $\overline{X}$ or S^2 is the better estimator of λ?

Activity 12 *Comparing estimators of the parameter μ of a normal distribution*

For a normal distribution $N(\mu, \sigma^2)$, the mean is equal to μ. A normal distribution is symmetric, so its median is equal to its mean. Hence, given a sample from a normal distribution, both the sample mean $\overline{x}$, and the sample median m, might be used as an estimate of μ, so that the random

variables $\overline{X}$ (the sample mean) and M (the sample median) may both be used as estimators of μ. In this activity, you will use the animation **Estimators of the normal mean** to compare these estimators. This animation is similar to the animation **Estimators of the Poisson parameter** used in Activity 11, only with a different modelling distribution, and different estimators.

Here, M denotes the random variable version of the sample median.

- Open the **Estimators of the normal mean** animation.

The animation simulates taking a large number (N) of samples, each of size n, from a normal distribution with parameters μ and σ. The p.d.f. of the $N(\mu, \sigma^2)$ distribution is shown on the left-hand side of the animation. The controls for the simulation are below this. There are fields for the parameters μ and σ, the sample size n and the number of samples to be drawn N. The default settings of μ, σ, n, and N are 10, 2, 10 and 1000, respectively.

For each individual sample (of size n), the animation calculates the observed sample mean $\overline{x}$, and the observed sample median m. We therefore have a sample of N values $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N$ of $\overline{X}$, and N values $m_1, m_2, \ldots, m_N$ of M. When the simulation is run, unit-area histograms of these samples will be plotted on the right-hand side of the animation (initially these two graphs are empty): the histogram of the sample means (i.e. $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N$) will appear in the top right-hand graph, while the histogram of the sample medians (i.e. $m_1, m_2, \ldots, m_N$) will appear in the bottom right-hand graph.

The mean and variance of the sample means (i.e. the mean and variance of $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N$) are given beneath the histogram of the sample means, while the mean and variance of the sample medians (i.e. the mean and variance of $m_1, m_2, \ldots, m_N$) are given beneath the histogram of the sample medians. As with the previous animation, we would like to compare the variances of the two estimators $\overline{X}$ and M, and so the animation calculates:

$$\text{Ratio of variances} = \frac{\text{Variance of the sample medians}}{\text{Variance of the sample means}}.$$

If this ratio is 1, this would suggest that the variances of the two estimators are the same. However, if this ratio is greater than 1, this would suggest that the variance of M is larger than the variance of $\overline{X}$, while if the ratio is less than 1, this would suggest that the variance of M is smaller than the variance of $\overline{X}$. The ratio of variances is given underneath the plot of the normal p.d.f. on the left-hand side.

(a) First you will consider whether the two estimators are unbiased.

- Click on the **Take sample** button to run the simulation.
- Run several more simulations (by clicking on the **Take sample** button) using different values of μ and σ.

Do you think that both $\overline{X}$ and M are unbiased estimators of μ?

(b) In order to decide which of the estimators $\overline{X}$ and M is the better estimator, we need to compare their variances.

- Set the value of the $\boldsymbol{\mu}$ field to be `10` and the $\boldsymbol{\sigma}$ field to be `2` (these were the default values).
- Set the **Sample size, *n*** field to be `2`, so that each sample has just two observations.
- Set the **Number of samples, *N*** field to a large value. (For most computers, 10 000 is a reasonable choice, but use a smaller value if the animation runs too slowly for you.)
- Click on the **Take sample** button to run the simulation.

What is the value of the ratio of variances when $n = 2$? Can you explain why the ratio takes this value?

(c) Now you will investigate what happens to the ratio of variances as the sample size n increases.

- Using the default values $\mu = 10$ and $\sigma = 2$, run simulations (by clicking on the **Take sample** button) varying the sample size n so that you can complete Table 4 below. (You found the value of the ratio of variances when $n = 2$ in part (b).)

Table 4 Observed ratio of variances for normal distribution samples of size n

n	2	3	5	10	20	50	100
Ratio of variances							

What happens to the ratio of variances as n increases?

(d) Repeat part (c) for different values of μ and σ of your choice. Does your conclusion from part (c) change for different values of μ and σ?

(e) Which do you think is the better estimator of μ, $\overline{X}$ or M?

It seems that the sample mean is a better estimator of the population mean of both a Poisson distribution and a normal distribution compared with the other natural estimators that we tried; these were the sample variance in the Poisson case and the sample median in the normal case. Be warned, however, that the best estimator of a parameter is not always based on the sample mean!

4 Interpreting confidence intervals

This chapter is associated with Section 2 of Unit 8.

So far you have met a confidence interval for the mean, μ, that is based on an approximation that is valid for large samples. In this chapter, you will use an animation to investigate the interpretation of such confidence intervals.

The large-sample confidence interval for the mean is valid for any underlying population distribution, and we start, in Activity 13, by considering confidence intervals when the underlying population distribution is the continuous uniform. In this activity, you will use the animation **Repeated experiments** to select random samples from the continuous uniform distribution, calculate confidence intervals for the mean, and verify that in large samples the proportion of intervals containing the population mean corresponds approximately to the confidence level. For example, a 90% confidence interval for the mean is calculated in such a way that, in repeated experiments with large samples, about 90% of the confidence intervals calculated in this way contain the population mean.

Activity 13 *Repeated experiments*

- Open the **Repeated experiments** animation.

The animation opens on a tabbed panel labelled **Uniform**. This panel allows you to simulate random samples from a continuous uniform distribution $U(a, b)$. A graph of the p.d.f. of the uniform distribution $U(a, b)$ is shown on the left. This uniform distribution will be taken to represent the population distribution. The population mean μ is also displayed. The default values of a and b are 0 and 1. With the default settings the population mean is 0.5.

You can change the values of a and b if you wish.

(a) Begin by obtaining 90% confidence intervals for the mean from random samples of size 100, as follows.

- Change the **Confidence level** field to 90.

- Ensure that the **Sample size, *n*** field is set to `100` (the default).
- Select **Single sample** to take samples one at a time.
- Click on the **Take samples** button.

The animation simulates a random sample of size 100 from $U(a,b)$, calculates the sample mean $\overline{x}$ and sample standard deviation s for this sample, and computes the 90% confidence limits (as given by Interval (3) in Unit 8):

1.645 is the 0.95-quantile of the standard normal distribution.

$$\left(\overline{x} - 1.645\,\frac{s}{\sqrt{100}},\, \overline{x} + 1.645\,\frac{s}{\sqrt{100}}\right) = \left(\overline{x} - 0.1645\,s,\, \overline{x} + 0.1645\,s\right).$$

The confidence interval is displayed graphically on the right-hand side as a horizontal line stretching from the lower 90% confidence limit to the upper 90% confidence limit. The line indicates whether or not this confidence interval contains the population mean: if the confidence interval contains μ then the line is black, but if it does not contain μ, then the line is red.

Underneath this graphical display, you are told how many confidence intervals contain μ, and the percentage that contain μ. Since you have just a single confidence interval so far, these should say `1` and `100.0`, respectively, if your simulated confidence interval contains μ (and is therefore black), otherwise they should say `0` and `0.0`, respectively.

- Take several more samples by clicking repeatedly on the **Take samples** button. Continue until about 50 samples have been selected.

If you would like to know the limits of any particular confidence interval, if you hover the mouse pointer over a confidence interval, its limits will be displayed under the plot.

- Click on the **Sort** button.

The confidence intervals on the right are now rearranged so that the interval with the lowest mean is at the top. This provides a different visualisation of how the confidence intervals vary between experiments.

What proportion of the confidence intervals produced contain the population mean now? The proportion should be close to 90%, but there is a very good chance that, due to random fluctuations, it will not be exactly equal to 90%. However, if you were to take more and more samples of size 100, then the proportion of confidence intervals containing the population mean should gradually settle down to a value close to 90%. Note that this value may not be exactly 90% because the confidence intervals are approximate.

Now take a large number of samples as follows.

- Deselect **Single sample** by clicking on its tick box.
- Ensure that the **Number of samples, *N*** field is set to 1000 (the default value).
- Now click on the **Take samples** button.

The graph on the right will show some of the 1000 90% confidence

intervals calculated. You can use the scroll bar on the right to look at the others. You may also find that using the **Sort** button is useful.

- Click on the **Take samples** button several times.

Each time the **Take samples** button is clicked, a new set of 1000 intervals is generated. On each occasion what proportion of the 1000 intervals contain the population mean?

(b) Now you will obtain 95% confidence intervals for the mean from random samples of size 100.

- Change the **Confidence level** field to `95`.
- Click on the **Take samples** button a few times.

What proportion of intervals contain the population mean?

(c) How do the widths of the 95% confidence intervals compare, on average, with the widths of the 90% confidence intervals?

In the next activity, you will investigate the repeated experiments interpretation of confidence intervals using data sampled from an exponential distribution.

Activity 14 *More repeated experiments*

- In the **Repeated experiments** animation, click on the tab labelled **Exponential**.

This panel allows you to simulate random samples from an exponential distribution, and the associated large-sample confidence intervals for the mean of that distribution. A graph of the p.d.f. of the exponential distribution with parameter λ is shown on the left. The default value of λ is 1, and the mean of this distribution is therefore also 1.

Obtain 95% confidence intervals for the mean from random samples of size 100, as follows.

- Set the **Confidence level** field to `95`, the **Sample size, n** field to `100`, and the **Number of samples, N** field to `1000` (the default values).
- If necessary, deselect **Single sample**.
- Click on the **Take samples** button a few times.

You should find that the proportion of confidence intervals that contain the population mean varies around a value close to 95%, in line with the repeated experiments interpretation.

Now explore the effect of sample size on the proportion of 95% confidence intervals that contain the population mean.

- Change the **Sample size, n** field to a small value, say `10`.
- Click on the **Take samples** button several times.

Is the proportion of confidence intervals containing the population mean close to 95%? How do you explain what you find?

So far, you have confirmed the validity of the repeated experiments interpretation of confidence intervals for large samples from uniform distributions and from exponential distributions. In fact, the interpretation is valid for large samples from virtually *any* distribution. If you wish, you can investigate this for yourself using the other tabbed panels. You should find that, whatever distribution you choose, for large samples, the proportion of confidence intervals that contain the population mean settles down to a value roughly equal to the confidence level.

5 Large-sample confidence intervals

This chapter is associated with Subsection 3.4 of Unit 8.

In this chapter, you will use Minitab to calculate many of the large-sample confidence intervals described in Sections 1 and 3 of Unit 8. You will first see how to find large-sample confidence intervals for population means. You will then learn how to find large-sample confidence intervals for proportions, and for differences between proportions.

Start Minitab now if it is not already open.

5.1 Confidence intervals for a population mean

You will begin by finding large-sample confidence intervals for the population mean. These are also called z-intervals in Unit 8. In Minitab, they are calculated using **1-Sample Z...** from the **Basic Statistics** submenu of the **Stat** menu. When using **1-Sample Z...**, you need to enter a value for the population standard deviation σ; if you do not know σ you should use the sample standard deviation s instead. So you need to calculate the sample standard deviation before you embark on calculating a confidence interval. In fact, it is always a good idea to begin a statistical analysis by plotting the data and calculating some summary statistics, in order to get a feel for them. You will be asked to do this in the following activities.

See Examples 2, 6 and 7 of Unit 8.

In Activity 15, the use of Minitab to find z-intervals is illustrated for the data on strength of glass fibres of length 1.5 cm that are introduced in Unit 8.

Activity 15 *Strength of glass fibres*

Data on strengths of glass fibres of length 1.5 cm are in the Minitab worksheet **glass-fibre.mtw**. Open this worksheet now. The single variable `Strength` contains the strengths for each of 63 glass fibres of length 1.5 cm.

(a) Calculate appropriate summary statistics for the variable `Strength`, including the mean and standard deviation, and obtain a histogram of the data. Recall that a simple way of obtaining the required summary statistics together with a histogram (or other graph) is via **Stat** > **Basic Statistics** > **Display Descriptive Statistics. . . .** One important reason for looking at a histogram is to check that the mean provides a sensible summary measure of location.

For example, if half the glass fibres had strength 1.25, say, and the other half had strength 1.75, say, then we could of course calculate the mean, but would it be a useful summary?

(b) In this part of the activity, you are going to obtain 90% and 95% confidence intervals for the mean strength of glass fibres. Before you do, think for a moment about which of the confidence intervals you would expect to be the wider. Then obtain the confidence intervals, as follows.

- Choose **Stat** > **Basic Statistics** > **1-Sample Z. . . .** The **One-Sample Z for the Mean** dialogue box will open.
- The drop-down list at the top of the **One-Sample Z for the Mean** dialogue box has two options: **One or more samples, each in a column** and **Summarized data**. These options correspond to two different ways of entering the data. Here the data are in a single column of the worksheet, so select **One or more samples, each in a column** from the drop-down list.
- Enter `Strength` in the field underneath the drop-down list.
- Enter the sample standard deviation calculated in part (a) in the **Known standard deviation** field.
- Do not select **Perform hypothesis test**. This should be selected only if you want to carry out a test. (This will be discussed in Chapter 7.)
- Click on the **Options. . .** button to open the **One-Sample Z: Options** dialogue box.
- You must enter the confidence level you require in the **Confidence level** field. For example, enter `90` for a 90% confidence interval. (Ignore the **Alternative hypothesis** drop-down list.)
- Click on **OK** to close this dialogue box, then click on **OK** and the confidence interval will be produced.

Minitab can calculate confidence intervals for more than one variable at a time by entering more than one variable name.

Which of the 90% and 95% confidence intervals for the mean strength of glass fibres actually is the wider? Does this agree with what you expected, and why?

5.2 Confidence intervals for a proportion

Approximate, large-sample, confidence intervals for a single proportion are calculated using **1 Proportion. . .** from the **Basic Statistics** submenu of the **Stat** menu.

Activity 16 *Cellulitis*

In Activity 13 of Unit 8, data are given on the recurrence or otherwise of leg cellulitis upon treatment with penicillin. Of 136 cellulitis sufferers so treated, 30 had a recurrence of leg cellulitis. In this activity you will use Minitab to calculate a 99% confidence interval for the proportion of patients whose leg cellulitis recurred when treated with penicillin. The data are in summarised form, rather than in a Minitab worksheet, and so any Minitab worksheet can be used. Calculate the confidence interval as follows.

- Choose **Stat > Basic Statistics > 1 Proportion. . . .** The **One-Sample Proportion** dialogue box will open.
- The drop-down list at the top contains two options for entering the data: **One or more samples, each in a column** and **Summarized data**. The **One or more samples, each in a column** option is used when the data consist of a set of outcomes of Bernoulli trials, coded 0 (failure) or 1 (success) stored in a column of a Minitab worksheet. The data on cellulitis recurrence are summarised, so click on **Summarized data**.

The unsummarised data would consist of 30 1s and 106 0s in some order.

- Minitab calls a 'successful' trial an event, so enter `30` in the **Number of events** field and `136` in the **Number of trials** field.
- Do not select **Perform hypothesis test**.
- Click on the **Options. . .** button to open the **One-Sample Proportion: Options** dialogue box.
- Enter `99` in the **Confidence level** field. Ignore the **Alternative hypothesis** drop-down list. The **Method** drop-down list has two options: **Exact** and **Normal approximation**. You should select **Normal approximation** to calculate a confidence interval based on the normal approximation derived using the Central Limit Theorem (as used in Unit 8).
- Finally, click on **OK** to close the dialogue box, then click on **OK** to produce the confidence interval.

Activities 17 and 18 will give you further practice at using Minitab to calculate large-sample confidence intervals for proportions.

Activity 17 *Snoring*

Norton, P.G. and Dunn, E.V. (1985) 'Snoring as a risk factor for disease: an epidemiological survey', *British Medical Journal*, vol. 291, no. 6496, pp. 630–2.

It has been suggested that snoring might increase the risk of heart disease. To investigate this, data were collected on the frequency of snoring among a sample of people with heart disease and a sample of people without heart disease. The Minitab worksheet **snoring.mtw** contains the data. Open the worksheet now.

The data are in the form of a 2×4 table. The first row relates to people with heart disease, the second row to people without heart disease. In this activity, you will use the data on people without heart disease, of which there are 2374 in total. The four columns classify the data by snoring frequency: they are named `Never`, `Occasionally`, `Often` and `Always`.

(a) Calculate an approximate large-sample 95% confidence interval for the proportion of people without heart disease who always snore.

(b) Calculate an approximate large-sample 95% confidence interval for the proportion of people without heart disease who never snore.

(c) From the results you obtained in part (b), deduce an approximate 95% confidence interval for the proportion of people without heart disease who snore at least occasionally.

Stat > Basic Statistics > 1 Proportion...

In Activity 18, you will learn how to use Minitab to calculate confidence intervals from data that have not been summarised. You will also learn how to use Minitab to *recode* one variable to create another.

Activity 18 *Industrial accidents*

The Minitab worksheet **accidents.mtw** contains data on the number of accidents suffered by each of 414 machinists over a period of time. Note that these data are of historical rather than contemporary relevance, referring to industrial conditions around 100 years ago. The aim of this activity is to estimate the proportion of machinists who experience at least one accident. Open the worksheet now.

Greenwood, M. and Yule, G.U. (1920) 'An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents', *Journal of the Royal Statistical Society*, vol. 83, no. 2, pp. 255–79.

(a) The data consist of one column named `Accidents`, containing the accident frequencies for the 414 machinists. Obtain the distribution of the number of accidents suffered by each worker, as follows.

- Select **Stat > Tables > Tally Individual Variables...**.
- In the **Tally Individual Variables** dialogue box, enter `Accidents` in the **Variables** field and check that the **Counts** option is selected. (**Counts** should be selected by default; the other options should not be selected.)
- Click on **OK**.

A table containing the numbers of accidents and the frequencies of each number will be displayed in the Session window. You should find that 8 is the maximum number of accidents suffered by any worker.

(b) Now create a second variable, `Accidents2`, which takes the value 0 for workers who suffered no accidents and 1 for workers who suffered one or more accidents, as follows.

- Select **Data > Code > Numeric to Numeric...**.
- In the **Code: Numeric to Numeric** dialogue box, enter the variable `Accidents` in the **Code data from columns** field.

- Enter the new variable name `Accidents2` in the **Store coded data in columns** field.
- In the first of the fields labelled **Original values (eg, 1:4 12)** enter `1:8` and in the corresponding **New** field enter `1`. This recodes all values in `Accidents` between 1 and 8 (including 1 and 8) as 1s in `Accidents2`; all other values (in this case, 0s) are left unaltered.
- Click on **OK** and the new column of values (of 1s and 0s) will be added to the worksheet.

Using the variable `Accidents2`, estimate the proportion of workers who experience one or more accidents, and obtain an approximate 99% confidence interval for this proportion, as follows.

Stat > **Basic Statistics** > **1 Proportion...**

- Obtain the **One-Sample Proportion** dialogue box.
- Select **One or more samples, each in a column** from the drop-down list at the top, and enter `Accidents2` in the field underneath it.
- Click on the **Options...** button to open the **One-Sample Proportion: Options** dialogue box.
- In the **One-Sample Proportion: Options** dialogue box, enter the confidence level required and select the option **Normal approximation** from the **Method** drop-down list.
- Click on **OK** to close the dialogue box, then click on **OK** to obtain the results.

The output follows the same layout as when you use the **Summarized data** option, and thus includes the estimated proportion as well as the confidence interval.

5.3 Confidence intervals for the difference between two proportions

Large-sample confidence intervals for the difference between two proportions are calculated in Minitab using **2 Proportions...** from the **Basic Statistics** submenu of the **Stat** menu.

Activity 19 *Snoring and heart disease*

You used this worksheet in Activity 17. The data were described there.

Open the worksheet **snoring.mtw**, or make it the active worksheet if it is already open. There were 2374 people without heart disease of whom 1355 never snore, and 110 people with heart disease in the sample of whom 24 never snore.

(a) Using your calculator with these data, estimate the difference between the proportion of people without heart disease who never snore and the proportion with heart disease who never snore.

(b) Calculate a 95% confidence interval for the difference between these proportions, as follows.

- Select **Stat** > **Basic Statistics** > **2 Proportions. . . .** The **Two-Sample Proportion** dialogue box will open.
- The drop-down list at the top of the **Two-Sample Proportion** dialogue box contains three options for entering the data: **Both samples are in one column**, **Each sample is in its own column**, and **Summarized data**. The snoring data are summarised, so select **Summarized data**.
- Enter the data for the two samples in the fields labelled **Number of events** and **Number of trials**.
- Click on the **Options. . .** button to open the **Two-Sample Proportion: Options** dialogue box.
- Make sure that the confidence level is set to `95` and that **Estimate the proportions separately** is selected in the **Test method** drop-down list. Ignore both the **Hypothesized difference** and **Alternative hypothesis** fields.
- Now click on **OK** to close the dialogue box, then click on **OK** to obtain the confidence interval.

The Minitab output includes the proportions in each group, the difference between the proportions (which you calculated in part (a)), and an approximate confidence interval for the difference. Ignore the output relating to tests.

What might this confidence interval suggest about snoring in people with and without heart disease?

In Activity 20 you will use **2 Proportions. . .** to obtain a large-sample confidence interval for the difference between two proportions when the data are not summarised.

Activity 20 *Hepatitis and sewerage workers*

The data in the worksheet **sewer.mtw** were collected for a study of the occupational risk to sewerage workers of hepatitis A infection through contact with raw sewage. Open this worksheet now.

Data provided by C.P. Farrington, The Open University.

The worksheet contains data on 228 sewerage workers. There are four variables: `Immunity`, `Age`, `Exposure`, `Children`. In this activity, you will be concerned only with the variables `Immunity` and `Children`. The variable `Immunity` takes the value 0 if the worker had not been infected with hepatitis A by the time of the survey, and 1 if he or she had. The variable `Children` takes the value 1 if the worker had no children, and 2 if he or she had one or more children. The aim of the activity is to investigate the association, if any, between infection rates and numbers of children.

Note that the data have not been summarised, and that the infection data for both groups of workers are contained in the single variable `Immunity`. Which group of workers (those with or without children) each data point belongs to is identified by the values 1 and 2 in the variable `Children`. These values are called *Sample IDs* in Minitab.

The solution will consider only the difference this way round.

(a) Estimate the difference between the proportion infected among sewerage workers without children and the proportion infected among sewerage workers with children, together with a 95% confidence interval for this difference, as follows.

Stat > **Basic Statistics** > **2 Proportions...**

- Obtain the **Two-Sample Proportion** dialogue box.
- Select **Both samples are in one column** from the drop-down list at the top of the **Two-Sample Proportion** dialogue box.
- Enter `Immunity` in the **Samples** field, and `Children` in the **Sample IDs** field.
- In the **Two-Sample Proportion: Options** dialogue box, make sure that the confidence level is set to `95` and that **Estimate the proportions separately** is selected from the **Test method** drop-down list.
- Finally click on **OK** to close the dialogue box, then click on **OK** to obtain the results.

(b) What might this confidence interval suggest about the risk of hepatitis A infection in this population?

6 Confidence intervals for normal means

This chapter is associated with Subsection 4.5 of Unit 8.

When it is reasonable to assume that the data are sampled from a normal distribution, then exact confidence intervals, known as t-intervals, may be used. In this chapter, you will see how to use Minitab to calculate t-intervals. This is done using two of the options in the **Basic Statistics** submenu of the **Stat** menu: these are **1-Sample t...** and **2-Sample t...**. In Subsection 6.1, you will calculate t-intervals for a single sample; in Subsection 6.2, you will calculate t-intervals for situations where you have two independent samples of observations.

6.1 Single samples

In Activity 21, you will see how to use Minitab to obtain a t-interval for a normal mean.

Activity 21 *Heights of schoolgirls*

The Minitab worksheet **schoolgirls.mtw** contains the heights (in cm) and weights (in kg) of 30 eleven-year-old schoolgirls from Bradford. Open the worksheet now.

Data provided by A.T. Graham, The Open University.

(a) Obtain a normal probability plot for the variable `Height`. Verify that the variation in height is adequately modelled by a normal distribution.

Graph > **Probability Plot...**

(b) Obtain a 90% t-interval for the mean height of eleven-year-old schoolgirls in Bradford, as follows.

- Choose **Stat** > **Basic Statistics** > **1-Sample t...**. This opens the **One-Sample t for the Mean** dialogue box. Notice that there is only one difference between this dialogue box and the dialogue box for **One-Sample z for the Mean**: you do not have to enter a value for the population standard deviation in this dialogue box.
- Select **One or more samples, each in a column** from the drop-down list at the top, and enter `Height` in the field below.
- Check that **Perform hypothesis test** is not selected – a test is not required here.
- Click on the **Options...** button to open the **One-Sample t: Options** dialogue box.
- Enter the confidence level in the **Confidence level** field. (Ignore the **Alternative hypothesis** drop-down list.)
- Click on **OK** to close the dialogue box, then click on **OK** to obtain the confidence interval.

6.2 Two independent samples

For the difference between two population means, t-intervals are based on the sampling distribution of the difference of two sample means. The validity of the calculations depends on the following assumptions.

- Each sample is drawn from a normal distribution.
- The two samples are independent.

Slightly different procedures are used according to whether the variances of the two populations can be assumed to be equal. In Unit 8, the following rule of thumb was given: if the larger variance divided by the smaller variance is less than three, then you can assume that the population variances are equal.

Equality of variance or otherwise is important for the following reason. If the two population variances are equal, then the resulting confidence intervals are exact. If the variances are not equal, then an approximate procedure must be used. In this case, Minitab uses Welch's t-interval, as

briefly described in Subsection 4.5 of Unit 8. In general, this produces slightly wider confidence intervals than those obtained if you assume the population variances are equal.

In Minitab, both approaches may be implemented using **2-Sample t...** from the **Basic Statistics** submenu of the **Stat** menu.

Activity 22 *Etruscan skulls*

Barnicot, N.A. and Brothwell, D.R. (1959) 'The evaluation of metrical data in the comparison of ancient and modern bones', in Wolstenholme, G.E.W. and O'Connor, C.M. (eds) *Medical Biology and Etruscan Origins*, Boston, Little, Brown & Co.

Extent of Etruscan civilisation

The origins of the Etruscan civilisation which dominated parts of Italy in the middle centuries of the first millennium BCE remain something of a mystery. One question is whether Etruscans were natives of Italy, or immigrants from elsewhere. In an anthropometric study, observations on the maximum head breadth (measured in mm) were taken on 84 skulls of Etruscan males and on 70 skulls of modern Italian males (presumed to be similar to those of native Italian males of the Etruscan period). These data may be used to compare the mean head breadths of Etruscan males and modern Italian males.

The data are in the Minitab worksheet **skulls.mtw**. Open this worksheet now. There are two variables, `Etruscans` and `Italians`.

(a) Calculate the difference between the mean skull breadth of Etruscans and that of Italians. Calculate the sample standard deviations for the two variables. Check for equality of variance using the rule of thumb given above.

(b) Check whether the assumption of normality is reasonable for the two variables using normal probability plots.

(c) Now construct a 95% t-interval for the difference between the mean skull breadth of Etruscans and that of Italians, assuming equal variances, as follows.

- Choose **Stat** > **Basic Statistics** > **2-Sample t...** to open the **Two-Sample t for the Mean** dialogue box.
- Minitab will accept data arranged in any of three formats: **Both samples are in one column**, **Each sample is in its own column** and **Summarized data**. In this case, the two samples are in different columns of the worksheet, so select **Each sample is in its own column** from the drop-down list at the top.

  You met these data formats when using **2 Proportions...**.
- Enter `Etruscans` in the **Sample 1** field and `Italians` in the **Sample 2** field.
- Click on the **Options...** button to open the **Two-Sample t: Options** dialogue box.
- Check that the confidence level is set at 95.0. (Ignore the **Hypothesized difference** and **Alternative hypothesis** fields.)
- The next step is to decide whether it is valid to assume that the two samples come from populations with equal variances: you did this in part (a). You should have found that the assumption is reasonable. So select **Assume equal variances**.

- Click on **OK** to close the dialogue box, then click on **OK** to obtain the confidence interval.

In addition to various summary statistics, the Minitab output gives the 95% t-interval and the pooled estimate of the common standard deviation used to calculate the confidence interval. The rest of the output can be ignored.

(d) The two-sample t-interval that you obtained in part (c) is exact, in the sense that it does not rely on any approximations. It does, however, rely on the assumption that the variances in the two populations are equal. Recalculate the confidence interval without making this assumption: that is, use **2-Sample t...** as above, but this time **Assume equal variances** should not be selected.

In Activity 22, you should have found that the confidence interval obtained without assuming equality of variances is very similar to the one obtained assuming equal variances; this is a consequence of the fact that the sample variances do not differ much in this case. It is preferable to use the **Assume equal variances** option when appropriate to do so, since the confidence interval it yields is exact. Note, however, that the default in Minitab is to do the calculation without assuming equal variances.

7 Testing hypotheses

This chapter is associated with Subsection 4.1 of Unit 9.

As you saw in Subsection 3.1 of Unit 9, the (one-sample) z-test can be used with a large sample ($n \geq 25$) for testing hypotheses concerning the mean. In Minitab, z-tests are carried out using **Stats** > **Basic Statistics** > **1-Sample Z...**, which you used when calculating z-intervals in Subsection 5.1.

As for calculating z-intervals, in order to carry out a z-test, Minitab needs to be given a value for the population standard deviation σ. In the common situation where σ is not known, the sample standard deviation s should be used in place of σ. In this case, you will need to calculate s (for instance, by using **Stat** > **Basic Statistics** > **Display Descriptive Statistics...**) before performing the test.

Start Minitab now if it is not already open. You will use Minitab to perform a z-test in the next activity.

Activity 23 *Driving practical test pass rate*

In addition to passing a driving theory test, UK drivers must pass a practical driving test to obtain a full driving licence. Like the driving theory test, there are a number of test centres across the UK where drivers can take the practical test. The Minitab worksheet **practical-test.mtw**

Minitab allows you to do tests on more than one variable at the same time, so if required you could enter more than one variable name.

Since you are not interested in a confidence interval here, it does not matter what value is in the **Confidence level** field.

contains the pass rates for 316 UK driving practical test centres over the period April 2014–March 2015. Open this worksheet now.

The worksheet contains 4 columns of data. The first column (`Centre`) lists the 316 test centres; the second and third columns (`Male` and `Female`) contain the pass rates (%) for males and females, respectively, at each centre, and the fourth column (`Total`) contains the overall pass rates (%) for each centre.

During the period April 2013–March 2014, the national driving practical test pass rate was 47.1%. In this activity you will use the data in the column `Total` to carry out a z-test of the hypotheses

$$H_0 : \mu = 47.1\%, \quad H_1 : \mu \neq 47.1\%,$$

where μ is the mean pass rate for the driving practical test nationally across all UK test centres during the period April 2014–March 2015. Carry out the z-test as follows.

- Use **Stat > Basic Statistics > Display Descriptive Statistics...** to obtain the sample standard deviation for `Total`.
- Choose **Stat > Basic Statistics > 1-Sample Z...** to open the **One-sample Z for the Mean** dialogue box.
- The data are contained in columns on the worksheet, so select **One or more samples, each in a column** from the drop-down list at the top of the **One-sample Z for the Mean** dialogue box.
- Enter `Total` in the field underneath.
- Enter the value `7.162`, the sample standard deviation for `Total` that you found earlier, in the **Known standard deviation** field.
- Select **Perform hypothesis test** and enter `47.1` in the **Hypothesized mean** field. (That is, enter 47.1 as the value of the mean specified in the null hypothesis.)
- Click on the **Options...** button to open the **One-Sample Z: Options** dialogue box.
- Check that the drop-down list for **Alternative hypothesis** is set to `Mean ≠ hypothesized mean`. This is the setting required for a two-sided test.
- Click on **OK** to close the dialogue box, then click on **OK** to carry out the test.

In the Session window, Minitab gives the following output:

```
Test of μ = 47.1 vs ≠ 47.1

The assumed standard deviation = 7.162

 Variable   N   Mean StDev SE Mean      95% CI          Z     P
 Total    316 49.629 7.162   0.403 (48.839, 50.419) 6.28 0.000
```

As can be seen from the output above, Minitab first gives some brief details of the test it has performed and then some results, including a confidence interval, the observed value of the z-test statistic (`Z`), and finally the p-value (`P`). Notice that the p-value is 0.000. This means that

the p-value rounded to three decimal places is 0.000 so that $p < 0.0005$, which is very small!

Since $p < 0.01$, there is strong evidence against the null hypothesis that the mean pass rate for the driving practical test nationally over the period April 2014–March 2015 is the same as the national pass rate over the same period in the previous year. Further, since the observed test statistic, z, is positive, or equivalently because the sample mean, 49.629, is greater than the hypothesised population mean 47.1, the test suggests that the mean pass rate nationally over the period April 2014–March 2015 is higher than the national pass rate over the same period in the previous year.

If the sample size is small and the underlying distribution can be assumed to be normal, **1-Sample Z...** can be used if the standard deviation is known. However, this situation is relatively uncommon. Typically, the standard deviation must be estimated using the data, in which case a (one-sample) t-test is appropriate. In Minitab this is done using **Stat** > **Basic Statistics** > **1-Sample t...**, which you used in Subsection 6.1 to find (one-sample) t-intervals. The **1-Sample t** dialogue box is the same as the **1-Sample Z** dialogue box except that you do not need to enter a value for the population standard deviation. Activity 24 will give you some practice at using **1-Sample t...** to carry out a t-test.

Activity 24 *Blueberries and systolic blood pressure: p-value*

Examples 6 and 12 of Unit 9 tested the hypotheses:

$$H_0 : \mu_S = 138 \text{ mm Hg}, \quad H_1 : \mu_S < 138 \text{ mm Hg},$$

where μ_S is the mean systolic blood pressure (BP) for menopausal women after taking 22 g of freeze-dried blueberry powder every day for 8 weeks. The sample mean and sample standard deviation systolic BP for 20 women taking 22 g of freeze-dried blueberry powder every day for 8 weeks were 131 mm Hg and 17 mm Hg, respectively. The systolic BP measurements were assumed to follow a normal distribution and so a t-test can be used to test the hypotheses. The observed value of the test statistic for this test was $t \simeq -1.841$, and the null distribution was $t(19)$.

Blueberries pre-powdering

In Example 12, the p-value associated with this t-test was given as being $p = 0.041$. In this activity you will use Minitab to confirm this value. The data are in summarised form, rather than in a Minitab worksheet, and so any Minitab worksheet can be used.

- In Minitab, choose **Stat** > **Basic Statistics** > **1-Sample t...** to open the **One-sample t for the Mean** dialogue box.
- Since the data are in summarised form and not stored as raw data in a column, select **Summarized data** from the drop-down list at the top of the **One-Sample t for the Mean** dialogue box.
- Enter the value `20` in the **Sample size** field, `131` in the **Sample mean** field, and `17` in the **Standard deviation** field.

- Select **Perform hypothesis test** and enter `138` in the **Hypothesized mean** field.
- Click on the **Options...** button to open the **One-Sample t: Options** dialogue box.
- Since $H_1 : \mu_S < 138$ mm Hg, select `Mean < hypothesized mean` from the **Alternative hypothesis** drop-down list.
- Click on **OK** to close the dialogue box, then click on **OK** to carry out the test.

Since you are not interested in a confidence interval, it does not matter what value is in the **Confidence level** field.

The Minitab output given in the Session window is similar to that given by **1-Sample Z...**, except that the test statistic is now denoted `T` (since we have a t-test this time). Notice that Minitab has not given us a confidence interval this time: this is because we have used a one-sided test (and you should ignore the `95% Upper Bound` given by Minitab). Observe that the value of the test statistic, `T`, is a rounded version of the value calculated in Example 6 and the p-value is indeed 0.041 as stated in Example 12.

When observations are paired (for example if two observations – 'before treatment' and 'after treatment' – are made on individual patients), then the differences could be calculated so that a (one-sample) z- or t-test could be performed. There is an alternative method that can be used in Minitab which doesn't require calculating the differences first. Both methods are illustrated for some paired data in Activity 25.

Activity 25 *Plant height*

The data are quoted in Fisher, R.A. (1942) *The Design of Experiments*, 3rd edn, London, Oliver and Boyd, p. 27.

Charles Darwin, one of the people who introduced the theory of evolution, measured the heights of fifteen pairs of plants of the species *Zea mays*, that is, maize or corn. Each plant had parents grown from the same seed – one plant in each pair was the offspring of a cross-fertilisation, the other of a self-fertilisation. The key feature of these data is that the two height measurements within each pair are not independent, since the two plants within a pair are genetically related. For each pair, the difference between the heights of the plants can be found.

In this activity, you will use Minitab to carry out a t-test of the null hypothesis that the underlying population mean difference, μ_D, is zero, against a two-sided alternative hypothesis, that is, of

$$H_0 : \mu_D = 0, \quad H_1 : \mu_D \neq 0.$$

The test assumes that a normal distribution is an adequate model for the differences. This assumption should be checked; and, in fact, a probability plot suggests that the normality assumption is plausible.

The data are in the worksheet **darwin.mtw**. Open this worksheet now. There are two variables, `Cross` and `Self`. (The heights are measured in eighths of an inch.)

Use **Calc > Calculator...**

(a) Create a variable called `Difference` containing the paired differences `Cross − Self`.

Use **1-Sample t...** to perform a t-test using the variable `Difference` to test the hypotheses concerning μ_D given above, and report your conclusion.

Stat > Basic Statistics > 1-Sample t...

(b) The same test can also be performed as follows using **Paired t...**.

- Choose **Stat > Basic Statistics > Paired t...**.
- When using **Paired t...** you do not need to create a variable containing the differences. In the **Paired t for the Mean** dialogue box, select **Each sample is in a column** from the drop-down list at the top, and enter `Cross` in the **Sample 1** field and `Self` in the **Sample 2** field.
- The hypotheses are set up using **Options...**. Click on the **Options...** button to open the **Paired t: Options** dialogue box. The value in the **Hypothesized difference** field should be the mean difference specified in the null hypothesis: `0.0` in this case. The **Alternative hypothesis** drop-down list should be set to `Difference` ≠ `hypothesized difference`.
- Click on **OK** to return to the main dialogue box, then click on **OK** to perform the test.

Note that, when using **1-Sample t...**, the hypothesised mean is entered in the main dialogue box.

Since you are not interested in a confidence interval, it does not matter what value is in the **Confidence level** field.

What do you conclude?

Notice that, when using **Paired t...** rather than **1-Sample t...**, you do not need to calculate the differences first. However, if you wanted to check the adequacy of a normal model for the differences by drawing a normal probability plot, then you would have to calculate the differences yourself since **Paired t...** (like **1-Sample t...**) does not have an option for drawing such a plot.

The final test covered in this chapter is testing a proportion with a large sample.

Activity 26 *Young adults living with parents*

Example 7, Activity 11 and Exercise 4 in Unit 9 all considered data from the Labour Force Survey published in January 2014 by the UK's Office for National Statistics. In Example 7, the proportion, p_W, of young adults aged between 20 and 34 years living with their parents in Wales was considered; in Activity 11, the respective proportion, p_{NI}, in Northern Ireland was considered; and Exercise 4 considered the respective proportion, p_S, in Scotland. The sample data for these regions are given in Table 5.

Table 5 Numbers of young adults living with parents

Region	Sample size	Number living with parents
Wales	254	68
Northern Ireland	307	111
Scotland	344	86

In this activity, you will use Minitab to calculate the p-value for testing the hypotheses:

$$H_0 : p_W = 0.25, \quad H_1 : p_W \neq 0.25.$$

The data are in summarised form, rather than in a Minitab worksheet, and so any Minitab worksheet can be used.

- In Minitab, choose **Stat** > **Basic Statistics** > **1 Proportion. . .** .
- Since the data are in summarised form and not stored as raw data in a column, select **Summarized data** from the drop-down list at the top of the **One-Sample Proportion** dialogue box.
- Enter `68` in the **Number of events** field (since 68 of the respondents in Wales are living with their parents) and `254` in the **Number of trials** field (because there were 254 people in the sample for Wales).
- Select **Perform hypothesis test**, and enter `0.25` in the **Hypothesized proportion** field (0.25 is the value of p_W specified in the null hypothesis).
- Click on the **Options. . .** button to open the **One-Sample Proportion: Options** dialogue box.
- In the **One-Sample Proportion: Options** dialogue box, check that the **Alternative hypothesis** drop-down list is set to `Proportion` ≠ `hypothesized proportion` (for a two-sided test).
- Because we have a large sample and are using the Central Limit Theorem normal approximation for our test, select **Normal approximation** from the **Method** drop-down list.
- Click on **OK** to close the dialogue box, then click on **OK** to perform the test.

Since you are not interested in a confidence interval here, it does not matter what value is in the **Confidence level** field.

In the Minitab output, the test statistic is denoted by `Z-Value` for this test: its value is 0.65, which is the same value of the test statistic as that calculated in Example 7 (correct to two decimal places). The p-value is 0.514, which is much larger than 0.1 and so there is little or no evidence against the null hypothesis that $p_W = 0.25$.

When testing a single proportion, Minitab uses a more complicated method for calculating p-values than that described in Unit 9. However, in many instances, the two methods lead to the same p-value.

Activity 27 *Northern Ireland and Scotland*

In this activity, use the notation and data given in Activity 26.

(a) Use Minitab to find the p-value to test the hypotheses:

$$H_0 : p_{NI} = 0.25, \quad H_1 : p_{NI} \neq 0.25.$$

Interpret the p-value.

(b) Use Minitab to find the p-value to test the hypotheses:

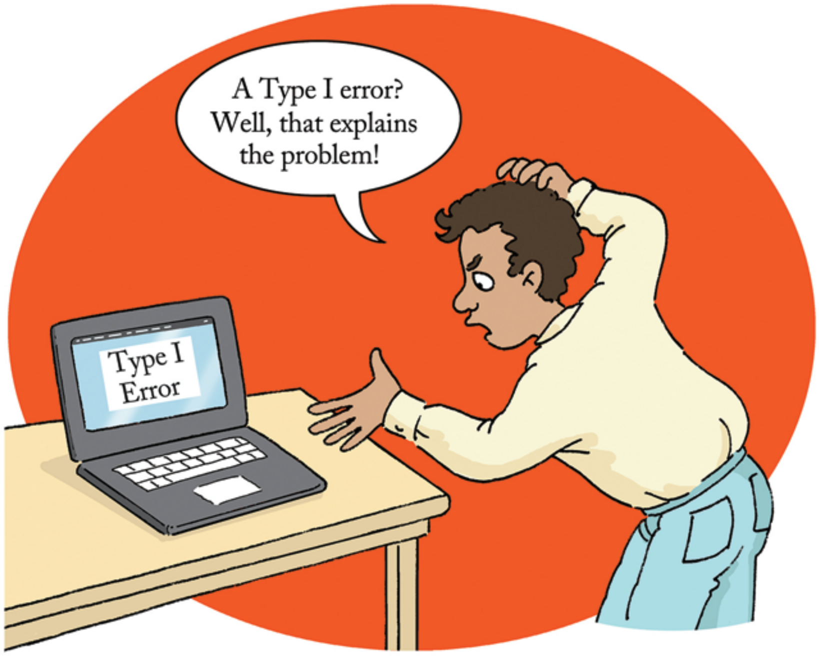
$$H_0 : p_S = 0.25, \quad H_1 : p_S \neq 0.25.$$

Interpret the p-value.

8 Type I and Type II errors

This chapter is associated with Section 5.1 of Unit 9.

In this chapter, you will use one of the M248 animations to explore Type I and Type II errors. You will start by considering Type I errors.



Activity 28 *Investigating P(Type I error)*

In this activity, you will simulate random samples from a normal distribution and use each sample to perform a t-test about the mean. You will verify that, in repeated sampling, the proportion of samples for which the null hypothesis is rejected, when in fact it is true, corresponds to the significance level of the test. This means that, for a test with significance level 0.05, for example, in a large number of repeated experiments where the null hypothesis is true, the null hypothesis will be (incorrectly) rejected in about 5% of the experiments. Each time the null hypothesis is incorrectly rejected, a Type I error is occurring, and so a Type I error should occur in about 5% of the experiments.

- Open the **Type I and Type II errors** animation.

A graph of the p.d.f. of the normal distribution $N(\mu, \sigma^2)$ is shown on the left-hand side of the animation. This normal distribution will be taken to represent the population distribution from which random samples are drawn. The values of the population mean μ and the population standard deviation σ are displayed underneath. The default values of the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ fields are `0` and `1`, respectively.

For each random sample simulated from $N(\mu, \sigma^2)$, the animation will carry out a t-test of the null hypothesis $H_0 : \mu = \mu_0$, for some specified value μ_0. The value μ_0 used in the t-test is set using the **Hypothesised mean, $\boldsymbol{\mu_0}$** field underneath the p.d.f. of the normal distribution. The default value for the **Hypothesised mean, $\boldsymbol{\mu_0}$** field is `0`. The null hypothesis will be true when the values of μ and μ_0 are the same, and so, under the default settings in the animation, the null hypothesis is true (both the $\boldsymbol{\mu}$ and **Hypothesised mean, $\boldsymbol{\mu_0}$** fields have default values of `0`).

The **Alternative** drop-down list has three options for choosing the alternative hypothesis: `less than` corresponding to $H_1 : \mu < \mu_0$, `not equal` corresponding to $H_1 : \mu \neq \mu_0$, and `greater than` corresponding to $H_1 : \mu > \mu_0$. The default option for the drop-down list is `not equal`, so that the default test is a 2-sided test.

The remaining fields on the left-hand side are **Significance level**, **Sample size, $\boldsymbol{n}$** and **Number of samples, $\boldsymbol{N}$**. The default values for these are `5`, `25` and `1000`, respectively, which correspond to simulating 1000 random samples, each of size 25, from $N(\mu, \sigma^2)$, and carrying out a t-test on each sample at the 0.05 significance level. Underneath these fields there is a **Single sample** tick box, which allows the animation to use individual samples one at a time: you will use this feature shortly.

For each simulated sample, the animation calculates the observed value of the test statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}},$$

where $\overline{x}$ and s are the sample mean and sample standard deviation of the simulated sample of size n from the population distribution. A graph of the p.d.f. of the null distribution of the test statistic T is shown on the right. This distribution is $t(n-1)$: so in the default case when the sample size is 25, the null distribution is a t-distribution with 24 degrees of freedom. Notice that this default graph has two shaded areas, one in each tail of the p.d.f. of the null distribution, since the default test is two-sided. These form the rejection region for the test, which together contain 5% of the probability in the null distribution (because the default significance level is 0.05). The boundaries of the rejection region for the default null distribution are therefore at the 0.025-quantile and the 0.975-quantile of $t(24)$, which are ± 2.064.

(a) We will start by simulating a single sample of size 25 from the normal distribution $N(0, 1)$.

- Select **Single sample** to allow you to take samples one at a time.
- Make sure that all other fields are set to their default values, so that the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ fields are set to `0` and `1`, respectively, the **Hypothesised mean, $\boldsymbol{\mu_0}$** field is set to `0`, the **Alternative** drop-down list is set to `not equal`, the **Significance level** field is set to `5`, and the **Sample size, $\boldsymbol{n}$** field is set to `25`. (The **Number of samples, $\boldsymbol{N}$** field can be ignored for now because **Single sample** has been selected.)

- Click on the **Take samples** button.

A single sample of size 25 will be simulated from $N(0,1)$. The individual values in the sample are marked with vertical lines on the p.d.f. of the normal population (on the left). The observed value t of the test statistic T is calculated using these 25 values, and marked on the p.d.f. of the null distribution (on the right). This value of t is displayed underneath the null distribution. The number of samples that have been taken so far (`1`) is also shown, together with the number and percentage of the samples that gave values of T in the rejection region: these will be `1` and `100.0`, respectively, if the test statistic lies in the rejection region, otherwise they will be `0` and `0.0`, respectively.

- Take several more samples, one at a time, by clicking repeatedly on the **Take samples** button.

You should find that, for most samples, the observed value of the test statistic is not in the rejection region, but that, occasionally, it is. This should come as no surprise. The population mean is 0 and you are testing the hypothesis that the population mean is 0. The null hypothesis is therefore true. So, since the significance level is 0.05, the null hypothesis should be rejected only 5% of the time, in the long run. As you take more and more samples, you should find that the percentage of values of T in the rejection region begins to settle down to about 5%. In other words, the probability of a Type I error is 0.05.

After you have taken several samples and observed the test statistic landing both in and out of the rejection region, stop taking samples one at a time, and take a larger number of samples as follows.

- Click on the **Single sample** tick box to deselect it. Ensure that the value in the **Number of samples, *N*** field is 1000.
- Click on the **Take samples** button.

Without showing each sample individually, 1000 samples will be taken, each of size 25. For each sample, the observed value of the test statistic is calculated and it is noted whether or not it falls in the rejection region. You should find that approximately 5% of the samples give values of T in the rejection region. That is, in repeated sampling, the proportion of samples for which the null hypothesis is rejected is approximately equal to the chosen significance level, and thus a Type I error occurs approximately 5% of the time.

(b) Now we will change the alternative hypothesis for the test from $H_1 : \mu \neq \mu_0$ to $H_1 : \mu > \mu_0$, so that the test is one-sided.

- Set the **Alternative** drop-down list to `greater than`.

Only positive values of the test statistic now provide support for the alternative hypothesis, so the rejection region consists entirely of points in the right-hand tail of the null distribution. However, the total probability in the rejection region is still 0.05 (the significance level), so the boundary of the rejection region is closer to zero than it was for the corresponding two-sided test. It is now the 0.95-quantile of $t(24)$, which is 1.711.

- Click on the **Take samples** button.

You should again find that approximately 5% of the samples give values of T in the rejection region.

- Now change the **Alternative** drop-down list to `less than`.

Only negative values of the test statistic provide support for the alternative hypothesis this time, and so the rejection region consists entirely of points in the left-hand tail of the null distribution. Since the significance level is still 0.05, the boundary of the rejection region is now -1.711.

- Click on the **Take samples** button.

Again, you should find that approximately 5% of the samples give values of T in the rejection region.

(c) You will now see what will happen to the rejection region if you change the significance level from 0.05 to 0.1, that is, from 5% to 10% in percentage terms.

- Click on the **Reset** button to restore the default settings.
- Type `10` in the **Significance level** field, then press **Enter** so that the graph of the null distribution is updated.

You should have noticed that the rejection region becomes larger when the significance level is increased from 5% to 10%, as its boundaries move in towards 0. This is because the total probability in the rejection region increases to 0.1. The boundaries of the region are the 0.05-quantile and the 0.95-quantile of $t(24)$, which are ± 1.711.

- Now click on the **Take samples** button.

You should find that about 10% of the samples give observed values of T in the rejection region.

Try several other values for the significance level. Whatever value you use for the significance level, you should find that the percentage of samples which give observed values of T in the rejection region is approximately the same as the significance level.

(d) Now you will investigate what happens when the sample size is changed.

- Click on the **Reset** button to restore the default settings.
- Change **Sample size, *n*** to a value of your choice and press **Enter**.
- Click on the **Take samples** button.
- Try several different sample sizes, including some very small sizes.

When the sample size is changed, the shape of the graph of the null distribution changes (because the distribution is $t(n-1)$, and the degrees of freedom change with the sample size). However, except for very small sample sizes, it is not easy to see the change. This is because, for moderate or large numbers of degrees of freedom, the p.d.f. of a t-distribution looks very similar to that of a standard

normal distribution. When the number of degrees of freedom is very small, the p.d.f. of a t-distribution is noticeably different from that of a standard normal distribution.

The value $n = 1$ is not allowed, and so if you type the value 1 into the **Sample size, $\boldsymbol{n}$** field, the animation will change the value to the smallest value allowed, 2.

Nevertheless, whatever the sample size, you should have found that the percentage of samples that give values of T in the rejection region is close to the significance level. The reason is that, for normally distributed data, the exact null distribution of the t-test statistic is $t(n-1)$.

So far, you have explored what happens when the null hypothesis is true. What happens when it is false? You are asked to investigate this in Activities 29 and 30.

Activity 29 *What happens when the null hypothesis is false?*

Now suppose that you are testing the null hypothesis $H_0 : \mu = 0$, but that the true population mean μ is not 0, but is 0.5. In this case, you would expect sample means to tend to be greater than 0, and values of the test statistic to be in the upper (right-hand) tail of the null distribution quite a lot of the time.

So **Hypothesised mean, $\boldsymbol{\mu_0}$** is still `0`.

- Click on the **Reset** button to restore the default settings.
- Change the value of the population mean $\boldsymbol{\mu}$ from `0` to `0.5`.
- Select **Single sample** and take a few samples one at a time.

You should notice that the observed value of t falls into the rejection region more frequently than it did when the null hypothesis was true, that is, when μ was equal to μ_0.

- Now deselect **Single sample** and take 1000 samples of size 25. What do you find?

Activity 30 *Changing the population mean*

What do you think would happen if you changed the value of the population mean μ from 0.5 to 1 and took 1000 samples of size 25? Try doing this and see if you are right.

If you have time, you might like to experiment further with changing the settings of the tests, to see the effect on the proportion of values of the test statistic that fall in the rejection region.

9 Power and sample size calculations

This chapter is associated with Subsection 5.4 of Unit 9.

In this chapter, you will learn how to use Minitab to calculate the power of a hypothesis test for tests discussed in Unit 9. You will also use Minitab to calculate the sample size that should be used if a test is to have a pre-specified power.

The dung beetle is the world's most powerful animal: it can carry up to 1100 times its own weight!

Power and sample size calculations are done in Minitab using **Power and Sample Size** from the **Stat** menu. The **Power and Sample Size** submenu includes the tests that you have already met (as well as others that you haven't met in M248!). Calculations for the situation discussed in Section 5 of Unit 9, where the underlying distribution is normal and its standard deviation is known, are performed using **1-Sample Z. . .** . The use of Minitab is illustrated in Activity 31 for this situation.

Start Minitab now if it is not already open.

Activity 31 *Using Minitab to calculate power*

Example 14 of Unit 9 considered testing the hypotheses:

$$H_0 : \mu = 5, \quad H_1 : \mu > 5,$$

using a sample of size 25 obtained from a normally distributed population $N(\mu, 100)$ and a significance level of 0.05. In that example, the power was 0.2578 when the true value of μ was assumed to be $\mu = \mu_0 + d = 7$, so that $d = 2$. You will now use Minitab to calculate this power.

You can use any Minitab worksheet.

Do the power calculation in Minitab as follows.

- Choose **Stat** > **Power and Sample Size** > **1-Sample Z. . .** to open the **Power and Sample Size for 1-Sample Z** dialogue box.
- Enter `25` in the **Sample sizes** field.
- Enter `2` in the **Differences** field.
- Leave the **Power values** field empty.
- Enter `10` in the **Standard deviation** field.
- Now click on the **Options. . .** button to open the **Power and Sample Size for 1-Sample Z: Options** dialogue box.
- In the **Power and Sample Size for 1-Sample Z: Options** dialogue box, select **Greater than** under **Alternative Hypothesis** (since we have $H_1 : \mu > 5$).
- Check that `0.05` is in the **Significance level** field. (The default value is 0.05.)
- Click on **OK** to close this dialogue box, and then click on **OK** in the main dialogue box.

The fields are labelled in the plural, because you can enter more than one value in each field. If you do, then Minitab will display the results in a table.

Minitab will report that the power of the test is $0.259511 \simeq 0.260$. Note that this is slightly different to the value given in Unit 9, which was

approximated slightly due to the normal tables being available only up to two decimal places.

A graph of the power function for a range of values of differences, d, is also displayed as in Figure 2. Notice how this power function plots the power against d rather than μ (as in Figure 15 in Unit 9): either d or μ can be used. (The graph will not be displayed if you deselect **Display power curve** in the **Power and Sample Size for 1-Sample Z: Graph** dialogue box.)

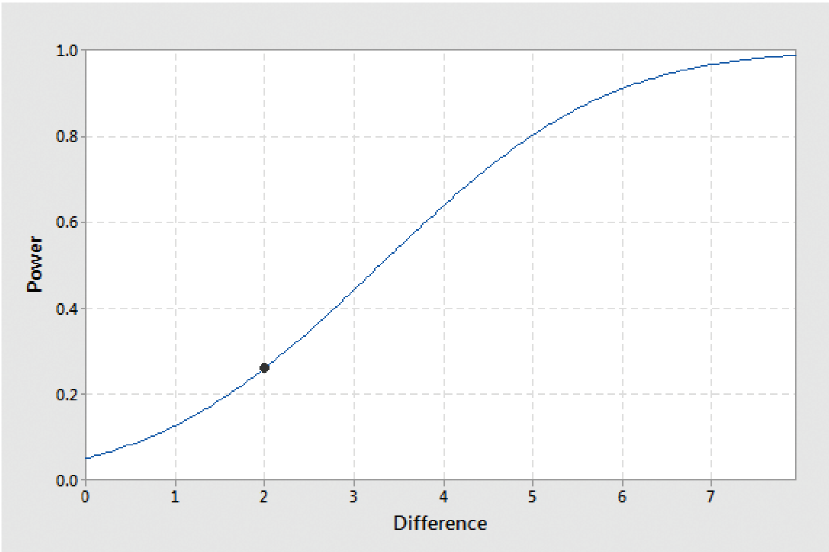


Figure 2 Power function for a one-sided test ($H_0 : \mu = 5$, $H_1 : \mu > 5$, significance level 0.05)

Activity 32 *Using Minitab to calculate sample size*

In this activity, you will use Minitab to calculate the sample size required in order for the power to be 0.9 in the testing scenario of Activity 31. The required sample size can be calculated as follows.

- Choose **Stat** > **Power and Sample Size** > **1-Sample Z. . . .**
- The **Sample sizes** field must be empty when a sample size calculation is required, so delete any value which is there.
- Ensure that `2` is in the **Differences** field.
- Enter `0.9` in the **Power values** field.
- Check that `10` is in the **Standard deviation** field, and that **Alternative hypothesis** is set to **Greater than** in the **Power and Sample Size for 1-Sample Z: Options** dialogue box.
- You do not need to change any of the other settings from those used in Activity 31. So click on **OK** to calculate the required sample size.

You can enter a list of power values here if a sample size calculation is required for more than one power value.

The power calculations in Activity 31 are not ideal because it was assumed that the population standard deviation, σ, was known to take the value 10. This assumption made it possible to perform a test that relied on the distribution of the sample mean being normal (a z-test). Typically, in practice, the population standard deviation is not assumed to be known, and a one-sample t-test is performed. For large sample sizes, this makes very little difference to the power calculations, but it can for small sample sizes. The details of how the computations are done for a t-test are more complicated than for a z-test, but when using Minitab, the calculations can be done just as easily for both tests.

Power and sample size calculations for t-tests are performed in Minitab using **Stat** > **Power and Sample Size** > **1-Sample t...**. Apart from its title, the dialogue box for calculating power and sample size is the same for **1-Sample t...** as for **1-Sample Z...**. In particular, it is still necessary to input a value for the population standard deviation. This might seem surprising, since the t-test does not use the population standard deviation. However, the power of a t-test (for a given difference d) depends on the population standard deviation σ, so some idea of the population standard deviation is needed in order to attempt to calculate the power of the test (or to calculate the sample size required for a given power). The sample standard deviation is the obvious value to use in situations where data have already been collected. (If data have not yet been collected, some idea of the population standard deviation is usually available from other studies.)

Activity 33 *Power for t-tests*

(a) Repeat the power calculations in Activity 31 using **Stat** > **Power and Sample Size** > **1-Sample t...**.

How does the power for the t-test compare with the power for the z-test?

(b) Now suppose that instead the sample size is 10 (rather than 25). How does the power for the t-test compare with the power for the z-test this time?

Activity 34 *Sample size for t-tests*

(a) Repeat the sample size calculations in Activity 32 using **Stat** > **Power and Sample Size** > **1-Sample t...**. How does the sample size for the t-test compare with the sample size for the z-test?

(b) Repeat part (a) for the easier situation where the difference that you wish to detect is much larger, say $d = 10$. How does the sample size for the t-test compare with the sample size for the z-test now? Can you account for what you observe?

10 The Wilcoxon signed rank test

This chapter is associated with Subsection 1.1 of Unit 10.

In Minitab, nonparametric tests are performed using **Nonparametrics** in the **Stat** menu. Carrying out a nonparametric test in Minitab is very similar to carrying out a parametric test. Indeed, on the whole, it is simpler because fewer options are available.

In this chapter, you will use Minitab to carry out the Wilcoxon signed rank test, which is used to test a null hypothesis about the location of a single sample of data (which might have arisen as differences between paired data values). You will start in Activity 35 by applying this test to data that you met in Unit 10.

Activity 35 *Corneal thickness*

The corneal thickness (in microns) was measured in the eyes of eight people, each of whom had one eye affected by glaucoma and one unaffected eye. These data were collected to investigate whether, on average, there is a difference between corneal thickness in the eye affected by glaucoma and the other eye.

These data were described in Activity 1 in Unit 10.

The data are in the Minitab worksheet **eyes.mtw**. Open the worksheet now.

There are two columns, called `Normal` and `Glaucomatous`, corresponding to the two eyes of each patient. In Unit 10, these data were analysed by calculating the difference between the corneal thicknesses of the eyes for each patient, and then testing the null hypothesis that these differences were drawn from a population with median 0, using the Wilcoxon signed rank test. A two-sided test was used. In this activity you will use Minitab to carry out the same test, as follows.

- Create a variable called `Difference` that contains the difference between the corneal thicknesses of the eyes (glaucomatous eye − normal eye) for each patient.

Calc > Calculator...

- Choose **Stat > Nonparametrics > 1-Sample Wilcoxon...** to open the **1-Sample Wilcoxon** dialogue box.
- In the **1-Sample Wilcoxon** dialogue box, enter `Difference` in the **Variables** field.
- The dialogue box allows you to choose between calculating a confidence interval for the population median and performing a hypothesis test. Select **Test median**.

Nonparametric confidence intervals are not described in M248.

- You are testing the null hypothesis that the population median is 0, so the **Test median** field should contain this value (which is the default). Check that it does so; if not, change it to `0`.
- A two-sided test is required, so check that the **Alternative** drop-down list is set to `not equal`.
- Finally, click on **OK**.

Check that you understand the Minitab output. What is the p-value for this test? What do you conclude?

In Activity 35, you saw that, for the Wilcoxon signed rank test, the calculation of the test statistic is the same as in Unit 10, but a different value is obtained for the p-value. The p-value given by Minitab is 0.310, which is different from the exact p-value of 0.344 reported in Example 3 of Unit 10, and is also different from the value of 0.271 calculated in Example 4 of Unit 10 using the normal approximation to the null distribution of the Wilcoxon signed rank test statistic.

In this case, since the effective sample size is only 7, the normal approximation is not very accurate.

The reason for this difference is that Minitab does not calculate the exact p-value for the Wilcoxon signed rank test. Instead, it uses the normal approximation described in Unit 10, but with an extra refinement. This refinement improves the approximation considerably so that it can be used for smaller sample sizes. If there are no tied ranks, the refined normal approximation provides a reasonable approximation to the exact null distribution for sample sizes as small as about 8. In Activity 35, the effective sample size was only 7 and there were two pairs of tied ranks, so you would not expect the normal approximation to be very good. But even in this case, Minitab's approximate p-value of 0.310 is a better approximation to the exact value of 0.344 than the value of 0.271 obtained by the basic normal approximation.

In Activity 35, you also saw that, as part of the output of the Wilcoxon signed rank test, Minitab produces an estimate of the population median, and that this estimate is not the sample median. Since the calculation of this estimate is not really part of the procedure for testing a hypothesis, the details of how it is calculated are not discussed here. In general terms, though, what is going on is as follows. As you saw in Unit 10, there is an assumption behind the Wilcoxon signed rank testing procedure that the underlying distribution is symmetric. If this assumption is justified, then it can be exploited to produce an estimator for the population median that has rather better statistical properties than the sample median, and it is this formula that Minitab uses to produce an estimate. (If the assumption of symmetry is not justified, then this estimator may well be worse than the sample median; but in such a case it would anyway not be valid to use the Wilcoxon signed rank test.)

Activities 36 and 37 will give you further practice at using Minitab to perform the Wilcoxon signed rank test, including using different choices for the hypotheses.

Activity 36 *Foetal movements*

The Minitab worksheet **movements.mtw** contains data on the differences between the percentage of time a foetus was moving before a test procedure and the percentage of time it was moving afterwards. Open this worksheet now. There is one variable, `Difference`.

In Activity 4 of Unit 10, you tested the null hypothesis that these differences could have been drawn from a population with median 0, using the Wilcoxon signed rank test. A two-sided test was performed.

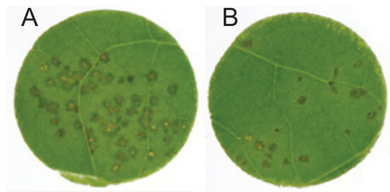
Carry out this test using Minitab.

Stat > Nonparametrics > 1-Sample Wilcoxon...

What p-value do you obtain? How does it compare with the exact p-value of 0.709 given in Unit 10? Would the difference, if any, change your conclusion concerning the result of the test?

Activity 37 *Viral lesions on leaves*

Two virus preparations (labelled A and B, respectively) were soaked into cheesecloth. The two viruses were then rubbed onto different halves of eight leaves. The number of local lesions appearing on each half was counted for each of the leaves.



Comparing lesions on leaves

Youden, W.J. and Beale, H.P. (1934) 'A statistical study of the local lesion method for estimating tobacco mosaic virus', *Contributions from Boyce Thompson Institute*, vol. 6, pp. 437–54.

The data are in the worksheet **lesions.mtw**. Open this worksheet now.

The columns `Alesions` and `Blesions` contain the numbers of local lesions appearing on the halves rubbed with virus A and on the halves rubbed with virus B, respectively. The data are paired: the two lesion counts for a particular leaf appear in the same row of the worksheet.

The question arises as to whether, on average, the two virus preparations tend to produce different numbers of lesions. Suppose that, in particular, the experimenters knew (from the way the preparations were produced) that preparation B was very unlikely to lead to more lesions, on average, than preparation A. Hence they wished to carry out a one-sided test with the alternative hypothesis that preparation A produces more lesions on average than preparation B.

(a) Write down the null and alternative hypotheses associated with this test.

(b) Create a variable containing the differences (A − B) between the numbers of local lesions on the two halves of each leaf. Produce a suitable plot to investigate whether these differences might reasonably be assumed to be a sample from a normal distribution. What do you conclude?

(c) Whether or not you considered a normal assumption to be reasonable in part (b), perform a one-sided Wilcoxon signed rank test on the differences (A − B). What do you conclude? Have you any reservations about the appropriateness of this test (as performed by Minitab) in this situation?

11 The Mann–Whitney test

This chapter is associated with Subsection 1.2 of Unit 10.

In this chapter, you will use Minitab to carry out another nonparametric test: the Mann-Whitney test. This test uses data from two independent samples, and may be used to investigate differences in location between the two populations involved. It is very straightforward to carry out the Mann–Whitney test using Minitab. This is first illustrated in Activity 38 for the dataset that was used in Examples 6 and 7 of Unit 10.

Activity 38 *Dopamine activity*

In a study into the causes of schizophrenia, 25 hospitalised patients with schizophrenia were treated with antipsychotic medication, and after a period of time were classified as psychotic or non-psychotic by hospital staff. A sample of cerebro-spinal fluid was taken from each patient and tested for dopamine β-hydroxylase enzyme activity.

The data, which are given in units of nmol/(ml)(h)/mg of protein, are in the worksheet **dopamine.mtw**. Open this worksheet now.

There are two columns: `Psychotic`, which gives the dopamine activities for the ten patients classified as psychotic, and `Non-psychotic`, which gives the dopamine activities for the other fifteen patients.

Analyse these data using a (two-sided) Mann–Whitney test, as follows.

- Choose **Stat** > **Nonparametrics** > **Mann-Whitney...** to open the **Mann-Whitney** dialogue box.
- The first two fields in the **Mann-Whitney** dialogue box, called **First Sample** and **Second Sample**, are for entering the variables containing the data for the two samples. In order to obtain output that matches the calculations in Unit 10, enter `Psychotic` in the **First Sample** field and `Non-psychotic` in the **Second Sample** field.
- The third field allows you to set the confidence level of the confidence interval that Minitab will calculate. Since we are not interested in this interval, you need not alter this field.
- We are considering a two-sided test here, and so the **Alternative** drop-down list should be set to `not equal`.
- Finally, click on **OK**.

(a) Can you interpret the output? (Hint: Minitab refers to the population medians as $\eta 1$ and $\eta 2$; η is a lower-case Greek letter, written ‘eta’ and pronounced ‘eater’.)

(b) Now repeat the Mann–Whitney test, but this time enter the samples the other way round (with `Non-psychotic` in the **First Sample** field). How does the output differ?

A place for Greek eaters?

As mentioned in the solution to Activity 38, the p-value calculated by Minitab in that activity (0.0025) is a little different to the exact p-value (0.0015) and that calculated in Example 7 of Unit 10 using a normal approximation (0.0022). The reason is that, as for the Wilcoxon signed rank test, Minitab uses the normal approximation given in Unit 10, but with the same kind of refinement. If there are no ties, the resulting p-values are reasonably accurate for sample sizes as small as 5 in each sample. If there are ties, then the approximation is (as usual) not so good. However, Minitab deals with ties by producing (using a somewhat complicated method) a better approximation to the p-value that allows for the ties. Thus the p-values reported by Minitab can generally be relied on unless the samples are very small, or there are many ties (or both).

The refinement usually improves the accuracy of the approximation. However, occasionally, as for the data in Activity 38, the approximation using the refinement is not quite as good!

This chapter concludes with two further activities to give you more practice at performing Mann–Whitney tests.

Activity 39 *Insect joint measurements*

Exercise 6 in Unit 8 concerned a research study on measurements on insect joints. More specifically, the data considered there and here are measurements made on the widths (in microns) of the first joints of the second tarsus for ten individuals from each of two species of the beetle *Chaetocnema*. These are *Chaetocnema concinna* (the mangold flea beetle) and *Chaetocnema picipes*. Interest lay in trying to discriminate between these two very similar-looking species of beetle on the basis of such measurements.

Individuals of the species of interest: *Chaetocnema concinna* (left) and *Chaetocnema picipes* (right)

In Unit 8, you produced a confidence interval for the difference between the population mean joint widths of the two species, having been told to assume that the distribution of joint widths in each population is normal. As so often with small samples of data, the assumption of normality is actually somewhat questionable and so here you will test whether the population difference between locations of the distributions of joint widths is the same or different in the two species, without making the normality assumption. The test will be two-sided because joints might be wider or narrower in one species compared with the other.

The data are in the worksheet **flea-beetle.mtw**. Open this worksheet now.

There are two columns, called `Concinna` and `Picipes`, containing respectively the widths of the first joints of the second tarsus for ten individuals from each of the two species of beetle named.

By performing a two-sided Mann–Whitney test, investigate whether there is a difference in location between the distributions of joint widths of the two species. What do you conclude?

Stat > Nonparametrics > Mann-Whitney...

Activity 40 *Oral socialisation and explanations of illness*

Whiting, J.W.M. and Child, I.L. (1953) *Child training and personality*, New Haven, CT, Yale University Press, p. 156.

A study was carried out of the relationship between customs related to illness and child-rearing practices in 39 non-literate societies.

On the basis of ethnographical reports, each of the societies was given a numerical rating for the degree of something called oral socialisation anxiety (OSA), which is a concept derived from psychoanalytic theory relating to child-rearing practice. (Socialisation refers to the acquiring of social skills.)

Each society was also placed into one of two groups. A judgement was made (by an independent set of judges) of whether oral explanations of illness were present in that society. The data therefore comprise one sample of OSA values for, as it turned out, 23 societies where oral explanations of illness were present, and a second sample of OSA values, of 16 societies where oral explanations of illness were absent. These data were collected to investigate a hypothesis that oral explanations of illness are more likely to be present in societies with high levels of OSA than in societies with low levels of OSA.

The data are in the worksheet **osa.mtw**. Open this worksheet now.

There are two columns, called `Present` and `Absent`, containing respectively the OSA scores for the societies where oral explanations of illness were judged to be present and the OSA scores for the societies where such explanations were judged to be absent. (There are many tied values in the data.)

If the hypothesis about the relationship between the presence of oral explanations of illness and oral socialisation anxiety were true, then you would expect OSA scores to tend to be higher in societies where oral explanations of illness were present than in societies where these explanations were absent. Thus, the alternative hypothesis is that the difference between the locations of the distributions of OSA scores where oral explanations of illness were present and where oral explanations of illness were absent is positive. On the other hand, if there were no relationship between the presence of oral explanations of illness and oral socialisation anxiety, then there would be no reason for the distribution of OSA scores to be different in societies where oral explanations of illness were present or absent. The null hypothesis is, therefore, that the difference between the locations of the distributions of OSA scores in the two groups of societies is zero.

By performing an appropriate one-sided Mann–Whitney test, investigate whether societies where oral explanations of illness are present tend to have higher OSA scores than societies where oral explanations of illness are absent. What do you conclude?

Exercises

Exercise 1 *Intelligence scores*

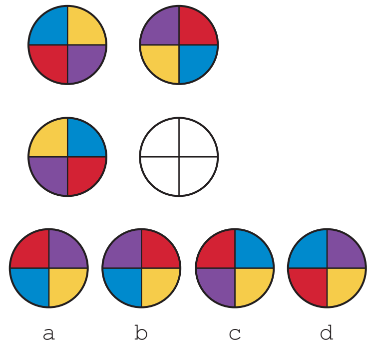
An IQ test is designed so that scores on the test are normally distributed with mean 100 and standard deviation 15.

(a) What proportion of the population tested record scores of 120 or more?

(b) What proportion of the population tested record scores between 110 and 130?

(c) Find the score such that only 15% of the population record that score or lower.

(d) Find the score such that only 5% of the population record that score or higher.



Which is the missing drawing? An example of the type of question often found in IQ tests.

Exercise 2 *Silver content of more coins*

Activity 8 investigated whether a normal distribution is a plausible model for the silver content for the first coinage (`Coin1`) of twelfth-century Byzantine coin data in the Minitab worksheet **coins.mtw**. Obtain normal probability plots for the other three coins, `Coin2`, `Coin3`, `Coin4`. Can you conclude that a normal distribution is a plausible model for each of these sets of coins as well?

Exercise 3 *Occupational risk to sewerage workers*

This exercise uses the same dataset as Activity 20, but there is no need to reopen the associated worksheet **sewer.mtw**. The data include various attributes of 228 sewerage workers. In particular: of the 50 sewerage workers with frequent exposure to raw sewage, 30 were found to have had hepatitis A; of the 178 sewerage workers with no or infrequent contact with sewage, 49 were found to have had hepatitis A.

(a) Calculate approximate 95% confidence intervals for the proportion with past infection in each of the two exposure groups from which the samples of workers were drawn.

(b) Calculate an approximate 95% confidence interval for the difference between the proportion of sewerage workers with frequent exposure to raw sewage who have had hepatitis A and the proportion with no or infrequent contact with raw sewage who have had hepatitis A.

(c) Interpret the confidence interval in part (b). Based on this confidence interval, does exposure to raw sewage constitute an occupational risk of hepatitis A infection?

Exercise 4 *Weights of schoolgirls*

The Minitab worksheet **schoolgirls.mtw**, previously considered in Activity 21, contains the heights (in cm) and weights (in kg) of 30 eleven-year-old schoolgirls from Bradford. A standardised measure of weight that is sometimes used in charts is weight divided by the number of centimetres above one metre height. In this exercise, you will investigate this measure using the data in **schoolgirls.mtw**.

(a) Create a new variable `Ratio` which contains the ratios of weight to height above 100 cm.

(b) Check whether it is reasonable to assume a normal model for the variation in the ratios.

(c) Calculate a 95% t-interval for the mean ratio. Note also the sample mean and sample standard deviation of the ratios.

(d) Calculate an approximate 95% confidence interval for the mean ratio. Compare this confidence interval with the confidence interval obtained in part (c).

Exercise 5 *Practical test pass rates: males and females*

The data for this exercise are in the Minitab worksheet **practical-test.mtw**. These are the same data that you used in Activity 23.

During the period April 2013–March 2014, the national driving practical test pass rate was 50.7% for males and 43.8% for females.

(a) Obtain the sample standard deviations for the pass rates for males and for females.

(b) Obtain the p-value for testing the hypotheses:

$$H_0 : \mu_M = 50.7\%, \quad H_1 : \mu_M \neq 50.7\%,$$

where μ_M is the mean driving practical test pass rate nationally for males only over the period April 2014–March 2015. What do you conclude?

(c) Obtain the p-value for testing the hypotheses:

$$H_0 : \mu_F = 43.8\%, \quad H_1 : \mu_F \neq 43.8\%,$$

where μ_F is the mean driving practical test pass rate nationally for females only over the period April 2014–March 2015. What do you conclude?

(d) In Activity 23, you performed a test of the hypotheses:

$$H_0 : \mu = 47.1\%, \quad H_1 : \mu \neq 47.1\%,$$

where μ is the mean driving practical test pass rate nationally for everyone over the period April 2014–March 2015. You found that $p < 0.01$ so that there was strong evidence against the null hypothesis, indications being that μ was higher than 47.1%. Bearing in mind that

in each of these three tests, the specific values being tested were the pass rates for the corresponding set of examinees in the period April 2013–March 2014, draw your results together and summarise your conclusions.

Exercise 6 *Heights of basketball players*

The data for this exercise are in the Minitab worksheet **basketball.mtw**. They comprise $n = 19$ heights of professional basketball players in the USA (in inches). You may assume that the distribution of the heights of basketball players can be modelled by a normal distribution.

Data taken from website of Department of Statistics, Florida State University.

(a) Use Minitab to obtain the mean and standard deviation of this sample of heights of basketball players.

(b) Obtain a 95% t-interval for the mean height of basketball players. Also obtain the p-value for testing the hypotheses:

$$H_0 : \mu = 82 \text{ inches}, \quad H_1 : \mu \neq 82 \text{ inches},$$

where μ is the mean height of basketball players. Hint: you should be able to use **Stat** > **Basic Statistics** > **1-Sample t...** to obtain both the confidence interval and the p-value at the same time.

(c) Interpret the p-value.

(d) What would have been the result of a hypothesis test at the 5% significance level? Explain how this result makes sense in relation to the 95% confidence interval obtained in part (b).

Exercise 7 *Sleep gain*

The pioneering statistician W.S. Gosset (who published his work under the pseudonym 'Student') analysed data on sleep gains (in hours) of ten patients after they had taken each of the drugs L-hyoscyamine hydrobromide and D-hyoscyamine hydrobromide. The data are in the worksheet **sleep-gain.mtw**, in columns named `Ltreatment` and `Dtreatment`.

'Student' (1908) 'The probable error of a mean', *Biometrika*, vol. 6, no. 1, pp. 1–25.

(a) Use a t-test to test the null hypothesis that the underlying mean sleep gain for patients who have taken L-hyoscyamine hydrobromide is zero, against the one-sided alternative hypothesis that the mean sleep gain is greater than zero. What do you conclude?

(b) Use a t-test to test the null hypothesis that the underlying mean sleep gain for patients who have taken D-hyoscyamine hydrobromide is zero, against the one-sided alternative hypothesis that the mean sleep gain is greater than zero. What do you conclude?

(c) Now use a one-sided Wilcoxon signed rank test to investigate again the question of whether each drug, considered on its own, leads to sleep gain on average. (You can request both analyses at the same time by entering both variable names in the **Variables** field in the **1-Sample Wilcoxon** dialogue box.) What do you conclude?

(d) The data in the worksheet are paired: the two values in each row are sleep gains for the same patient. Calculate the difference for each pair (`Ltreatment` − `Dtreatment`). Investigate whether a normal model is appropriate for these differences.

(e) Carry out an appropriate two-sided test to investigate whether the two substances differ in their effects on sleep.

Exercise 8 *Laurel and Hardy films*

Redfern, N. (2012) 'Shot length distributions in the short films of Laurel and Hardy, 1927 to 1933', *Cine Forum*, vol. 14, pp. 37–71.

Stan Laurel and Oliver Hardy were a famous comedy double act of the early twentieth century. They made many short comedy films, starting in the era of silent films and continuing into the era of sound cinema. Historians of film are interested in comparing aspects of silent and sound film making. Here, in particular, we will look at the lengths (in seconds) of the shots (i.e. the continuous footage between edits or cuts) in each type of film. The data we have are the median shot lengths in each of a number of silent Laurel and Hardy films and of Laurel and Hardy films with sound ('talkies').

Stan Laurel and Oliver Hardy

The data are in the worksheet **laurel-and-hardy.mtw**. There are two columns in the worksheet, with the self-explanatory names `Silent` and `Sound`.

By performing an appropriate nonparametric test, investigate whether the median shot lengths were the same in Laurel and Hardy's talkies as in their silent films. What do you conclude?

Solutions to activities

Solution to Activity 1

Diagrams showing the proportions and quantiles required are given in Examples 4 and 5 of Unit 6.

(a) If X is a random variable representing the chest measurement of a Scottish soldier, then the probability required is $P(37 \leq X \leq 42)$. This is equal to $F(42) - F(37)$, where F is the c.d.f. of X.

To find $F(42)$, in the **Normal Distribution** dialogue box, select **Cumulative probability**, enter `42` in the **Input constant** field, then click on **OK**. The value of $F(42) = P(X \leq 42)$ will be displayed in the Session window: this is 0.841345.

To find $F(37)$, in the **Normal Distribution** dialogue box, enter `37` in the **Input constant** field and click on **OK** (all other entries remain unchanged from the previous calculation): this gives $F(37) = P(X \leq 37) = 0.0668072$.

So, according to the model, the proportion of Scottish soldiers whose chest measurements were between 37 and 42 inches is

$$F(42) - F(37) = 0.841345 - 0.0668072 = 0.7745378 \simeq 0.7745.$$

(b) The proportion of Scottish soldiers whose chest measurements were greater than 43 inches is

$$1 - F(43) = 1 - 0.933193 = 0.066807 \simeq 0.0668.$$

(c) The chest measurements of only 2.5% of Scottish soldiers were below the 0.025-quantile of $N(40, 4)$.

In the **Normal Distribution** dialogue box, select **Inverse cumulative probability**, enter `0.025` in the **Input constant** field and click on **OK**. Minitab will display the value 36.0801. So, according to the model, the chest measurements of only 2.5% of Scottish soldiers were below about 36.1 inches.

(d) If 5% of soldiers had chest measurements above y, say, then 95% had chest measurements below y, so

$$y = q_{0.95} = 43.2897 \simeq 43.3 \text{ inches}.$$

Solution to Activity 2

To do this activity, use the facilities of Minitab in the same way as you did in Activity 1.

In the **Normal Distribution** dialogue box, enter `0` for the mean of the normal distribution and `1.658` for the standard deviation. Use **Cumulative probability** to answer parts (a) to (d), and **Inverse cumulative probability** to answer parts (e) to (g). Diagrams showing the probabilities and quantiles required are given in the solutions to Activities 6 and 7 of Unit 6.

$\sqrt{2.75} = 1.658$ to three decimal places.

(a) The probability required is $P(X > 0.5)$, which is given by

$$1 - F(0.5) = 1 - 0.618509 \simeq 0.3815.$$

(b) The probability required is $P(0 < X < 2)$, which is given by

$$F(2) - F(0) = 0.886144 - 0.5 \simeq 0.3861.$$

(c) Here the probability required is $P(-1 < X < 1)$, which is given by

$$F(1) - F(-1) = 0.726791 - 0.273209 \simeq 0.4536.$$

(d) The probability required is $P(|X| > 1.5)$, which is equal to either $2P(X > 1.5)$ or $2P(X < -1.5)$. This comes to $2 \times 0.182811 \simeq 0.3656$.

(e) The required value x is the 90% point of $N(0, 2.75)$, so $x = q_{0.9} \simeq 2.1248$.

(f) The required value b is the 97.5% point of $N(0, 2.75)$, so $b = q_{0.975} \simeq 3.2496$.

(g) The required value c is the 99.5% point of $N(0, 2.75)$, so $c = q_{0.995} \simeq 4.2707$.

Solution to Activity 3

Let X be a random variable representing the height in centimetres of a randomly selected elderly woman: $X \sim N(160, 6^2)$.

Calc > Probability Distributions > Normal...

In the **Normal Distribution** dialogue box, enter `160` for the mean of the normal distribution and `6` for the standard deviation. Use **Cumulative probability** to answer parts (a) and (b), and **Inverse cumulative probability** to answer parts (c) and (d).

(a) The proportion of elderly women taller than 166 cm is given by the shaded area in Figure 3. The probability required is

$$P(X > 166) = 1 - F(166) = 1 - 0.841345 \simeq 0.1587.$$

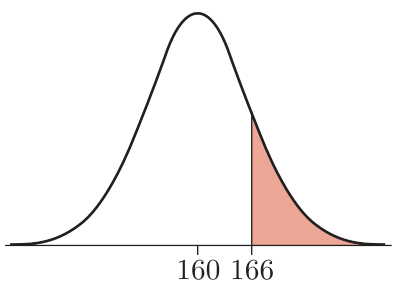


Figure 3

(b) The shaded area in Figure 4 represents the probability required. This probability is equal to

$$P(145 \leq X \leq 157) = F(157) - F(145) = 0.308538 - 0.0062097 \simeq 0.3023.$$

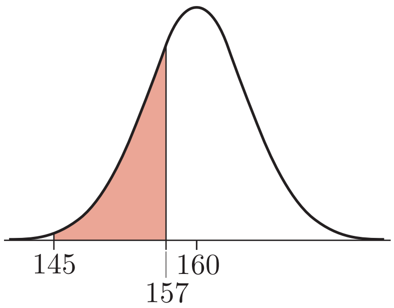


Figure 4

(c) The height required (in cm) is given by

$$q_{0.85} = 166.219 \simeq 166.2.$$

(d) The interquartile range (in cm) is

$$q_{0.75} - q_{0.25} = 164.047 - 155.953 = 8.094 \simeq 8.1.$$

Solution to Activity 4

Let X be a random variable representing the blood plasma nicotine level of a randomly selected smoker: $X \sim N(315, 131^2)$.

In the **Normal Distribution** dialogue box, enter `315` for the mean of the normal distribution and `131` for the standard deviation. Use **Cumulative probability** to answer parts (a) and (b), and **Inverse cumulative probability** to answer part (c).

Calc > Probability Distributions > Normal...

(a) The shaded area in Figure 5 represents the proportion of smokers with nicotine levels lower than 300. This is equal to

$$P(X < 300) = F(300) = 0.454419 \simeq 0.4544.$$

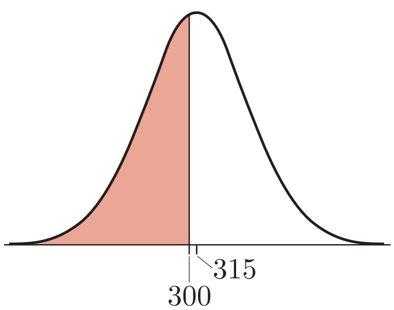


Figure 5

(b) The shaded area in Figure 6 (overleaf) represents the proportion of smokers with nicotine levels between 150 and 400. This is equal to

$$P(150 \le X \le 400) = F(400) - F(150) = 0.741784 - 0.103917 \simeq 0.6379.$$

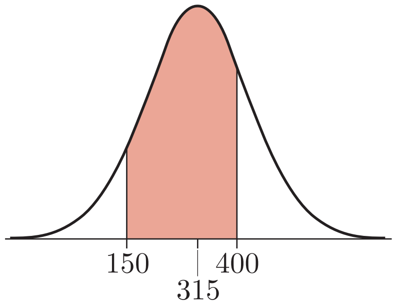


Figure 6

(c) The nicotine level required is the 80% point of $N(315, 131^2)$, which is

$$q_{0.8} = 425.252 \simeq 425.3.$$

Solution to Activity 5

(a) For $X \sim N(40, 2^2)$, the proportion of measurements more than three standard deviations from the mean is given by

$$\begin{aligned} P(X < 34) + P(X > 46) &= F(34) + (1 - F(46)) \\ &= 0.0013499 + (1 - 0.998650) \\ &\simeq 0.0027. \end{aligned}$$

(b) This time the probability required is

$$\begin{aligned} &P(X < 0 - 3 \times 1.658) + P(X > 0 + 3 \times 1.658) \\ &\qquad = F(-4.974) + (1 - F(4.974)). \end{aligned}$$

This is also equal to 0.0027 to four decimal places.

(c) In this case, the probability required is $P(X < 142) + P(X > 178)$, which comes to 0.0027 to four decimal places.

(d) In each case, the proportion of observations more than three standard deviations from the mean is approximately equal to 0.0027. In fact, the proportion of observations more than three standard deviations from the mean is equal to 0.0027 for *any* normal distribution, whatever the values of the mean μ and the standard deviation σ.

Solution to Activity 6

(a) For each pair of values of μ and σ that you tried, you should have found that

- the proportion of observations within one standard deviation of the mean is approximately $0.841345 - 0.158655 \simeq 0.6827$
- the proportion of observations within two standard deviations of the mean is approximately $0.977250 - 0.0227501 \simeq 0.9545$.

(b) Whatever value of k you chose, you should have found that the proportion of observations within k standard deviations of the mean does not depend on the values of the mean μ and the standard deviation σ.

(c) In general, for any $k > 0$, the proportion of observations within k standard deviations of the mean does not depend on the values of the

mean μ and the standard deviation σ. That is, for $X \sim N(\mu, \sigma^2)$, the value of the probability $P(\mu - k\sigma < X < \mu + k\sigma)$ depends only on the value of k and not on μ and σ.

Solution to Activity 9

The normal probability plot for the data on nicotine levels of smokers is shown in Figure 7.

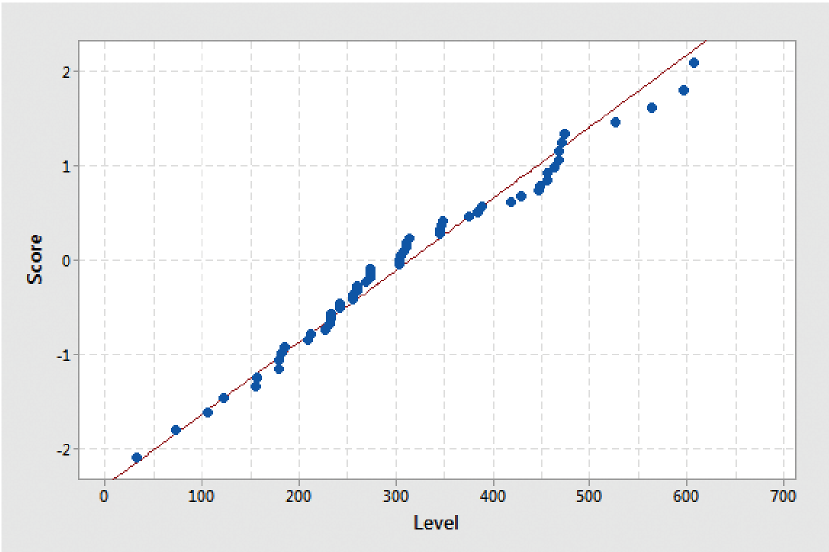


Figure 7 A normal probability plot for the nicotine levels

The points lie roughly along a straight line, suggesting that a normal model is plausible for the blood plasma nicotine levels of smokers.

Solution to Activity 10

The normal probability plot for the width-to-length ratios is shown in Figure 8.

Graph > Probability Plot...

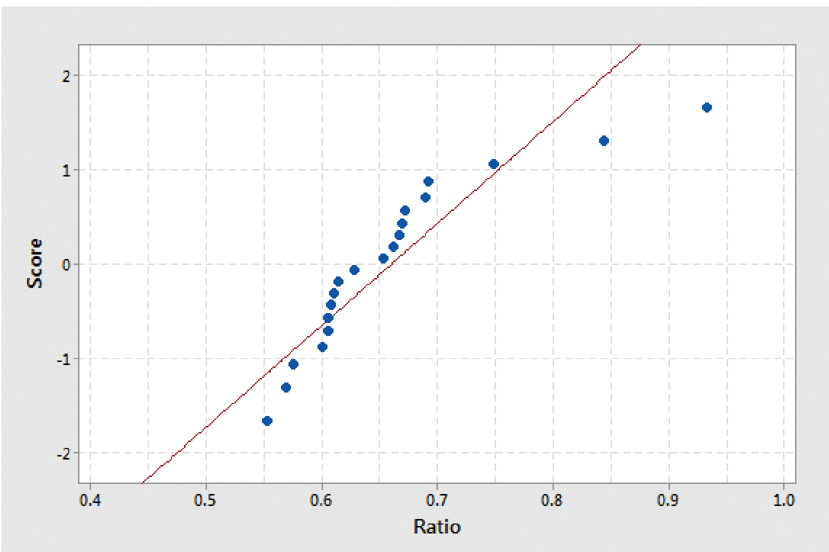


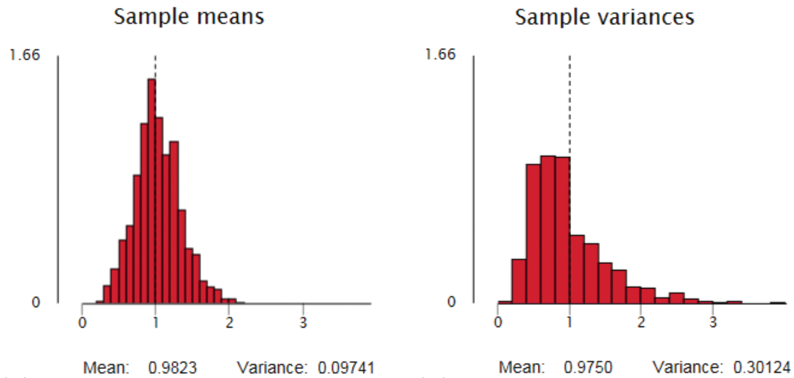
Figure 8 A normal probability plot for the width-to-length ratios

The points seem to lie roughly along a curve, not along a straight line. So it looks as though a normal model is not appropriate here.

Solution to Activity 11

The histograms that you produce are likely to look different to these because they are based on simulated random samples.

(a) The histogram for the sample means should have been roughly symmetric, while the histogram for the sample variances was probably clearly right-skew. Figure 9 shows histograms of the sample means and sample variances produced by the animation after taking 1000 samples, each of size 10, from Poisson(1).



(a) (b)

Figure 9 Histograms of (a) the sample means, and (b) the sample variances, for 1000 samples, each of size 10, from Poisson(1)

(b) You should have found that both the mean of the sample means, and the mean of the sample variances, are generally close to 1. Since the value of λ is 1, this suggests that both $\overline{X}$ and S^2 are unbiased estimators.

(c) The ratio of variances should generally have been around the value 3, so that the variance of the observed values of S^2 was generally roughly three times as large as the associated variance of the observed values of $\overline{X}$. So, since both estimators are unbiased, $\overline{X}$ is the preferable estimator when n is 10 and λ is 1.

(d) You should have found that the histogram for the sample means is slightly right-skew for $n = 5$ and $\lambda = 1$ but any skewness was probably not apparent for larger values of n or λ. This illustrates the Central Limit Theorem: as the sample size n increases, the distribution of $\overline{X}$ approaches a normal distribution. It is likely that all your histograms for the sample variances were right-skew. The skewness will have decreased as n increased and as λ increased, but skewness should have been apparent even for the largest values that you used. Figure 10 shows histograms of the sample means and sample variances produced by the animation after taking 10 000 samples, each of size 5, from Poisson(1), while Figure 11 shows histograms of the sample means and sample variances produced after taking 10 000 samples, each of size 50, from Poisson(27).

The histograms that you produce are likely to look different to these because they are based on simulated random samples.

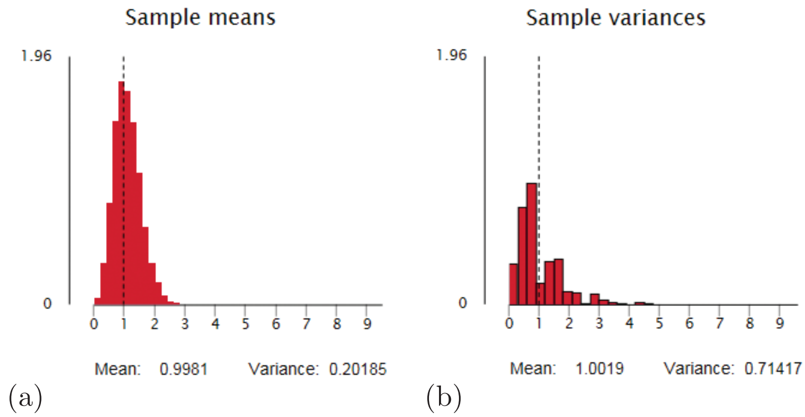


Figure 10 Histograms of (a) the sample means, and (b) the sample variances, for 10 000 samples, each of size 5, from Poisson(1)

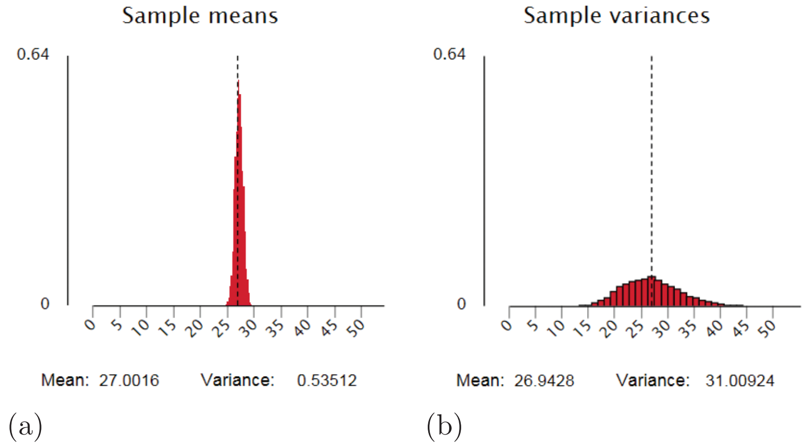


Figure 11 Histograms of (a) the sample means, and (b) the sample variances, for 10 000 samples, each of size 50, from Poisson(27)

The mean values you obtained for the sample means and the sample variances should have been close to the real value of λ in every simulation. You probably found that the means were sometimes just above λ and sometimes just below it. In fact, $\overline{X}$ and S^2 are unbiased estimators of λ for any values of λ and n.

You should have found that the ratio of the variances is always well above 1 (the smallest value should be about 3). The ratio decreases slightly as n increases, but increases quickly as λ increases. As the variance is always much smaller for $\overline{X}$ than for S^2, it seems clear that $\overline{X}$ is a much better estimator than S^2.

Solution to Activity 12

(a) You already know that $\overline{X}$ is an unbiased estimator of the population mean for any distribution, and so, in particular, $\overline{X}$ is an unbiased estimator of the normal mean. It should have come as no surprise to observe that the mean of the sample means is generally close to whichever values of μ you used.

You should have also made a similar observation concerning the mean of the sample medians: the mean of the sample medians is generally close to whichever value of μ you used, suggesting that M is also an unbiased estimator of μ.

(b) When $n = 2$, the ratio of the variances is 1. This is because, for a sample of size 2, both the sample mean and the sample median are calculated as the sum of the two observations in the sample divided by 2. Hence the histograms for the sample means and the sample medians will be identical, as will be their means and variances.

(c) You should have found that for all values of sample size n except 2, the ratio of variances is greater than 1. There is, however, no particular pattern to the ratio of variances as n increases.

(d) You should find that no matter what values are chosen for μ and σ, the ratio of variances is greater than 1 for values of n greater than 2, and is equal to 1 when n is 2.

(e) Since both estimators are unbiased, the one with the smaller variance will be the better estimator. So $\overline{X}$ seems to be better than M as an estimator of the parameter μ of a normal distribution since the ratio of variances is greater than 1 for all values of n except 2 (and equal to 1 when $n = 2$).

Solution to Activity 13

(a) You should have found that your results support the repeated experiments interpretation of confidence intervals: on average, the proportion of 90% confidence intervals containing the population mean is about 90%.

(b) The proportion of 95% confidence intervals containing the population mean is about 95%.

(c) You should have found that 95% confidence intervals are, on average, wider than 90% confidence intervals. In general, the higher the confidence level, the wider the interval.

Solution to Activity 14

You should have found that, for small sample sizes (of size 10, say) the proportion of 95% confidence intervals containing the population mean can be rather different from 95%. (In fact, it seems to be rather less than 95%.) For larger sample sizes (in the case of the exponential distribution, for sample sizes of 100 or more) the proportion of 95% confidence intervals containing the population mean is close to 95%. The conclusion from this activity is that the repeated experiments interpretation is valid for large samples from an exponential distribution, and the approximations improve as the sample size gets larger. This is because the assumptions on which the confidence intervals are based are valid for large sample sizes, but not for small ones.

Solution to Activity 15

(a) The mean of `Strength` is 1.5068 and the standard deviation is 0.3241. The histogram has a strong mode (around 1.6) and some suggestion of

skewness to the left, but there is no reason not to suggest that the mean provides a good summary of location.

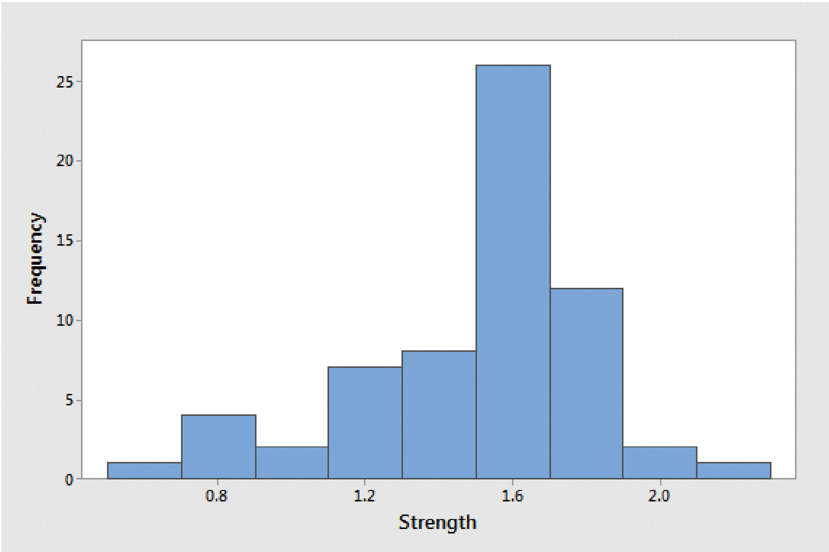


Figure 12 A histogram of strengths

(b) An approximate large-sample 90% confidence interval for the mean strength is $(1.4397, 1.5740)$; an approximate large-sample 95% confidence interval is $(1.4268, 1.5869)$. The 95% confidence interval is wider than the 90% confidence interval. This is because a higher confidence level means more confidence that the interval contains the population mean and so has to contain more values.

The 95% confidence interval was obtained in Example 7 of Unit 8.

Solution to Activity 16

The Minitab output includes the sample proportion $p = 30/136 \simeq 0.220588$ and the approximate 99% confidence interval $(0.129004, 0.312173)$. When reporting results, it is a good idea to round results which have been given to so many decimal places by Minitab. So in this case, we would perhaps report that the proportion 0.221 of leg cellulitis patients suffer a recurrence when treated with penicillin, with 99% confidence interval $(0.129, 0.312)$.

Solution to Activity 17

The confidence intervals required are calculated using the **Summarized data** option from the drop-down list at the top of the **One-Sample Proportion** dialogue box. To obtain the approximate confidence interval, remember to select the **Normal approximation** option from the **Method** drop-down list of the **One-Sample Proportion: Options** dialogue box.

(a) The number of trials is 2374, and the number of successes (events) is 224. The proportion who always snore is estimated to be 0.094, with approximate 95% confidence interval $(0.083, 0.106)$.

(b) The number of trials is 2374, and the number of successes is 1355. The proportion who never snore is estimated to be 0.571, with approximate 95% confidence interval $(0.551, 0.591)$.

(c) The proportion who snore at least occasionally (p_1, say) is equal to one minus the proportion who never snore (p_2). Thus $p_1 = 1 - p_2$, a decreasing transformation. The proportion who snore at least occasionally is estimated to be $1 - 0.571 = 0.429$. The confidence limits of an approximate 95% confidence interval for this proportion are obtained using Interval (8) of Unit 8 as

$$p_1^- = 1 - p_2^+ = 1 - 0.591 = 0.409$$

and

$$p_1^+ = 1 - p_2^- = 1 - 0.551 = 0.449.$$

Therefore, the approximate 95% confidence interval for p_1 is (0.409, 0.449). The point is that you do not need to calculate the confidence interval from scratch; it is simpler to calculate it directly from the one you obtained in part (b).

Solution to Activity 18

(a) The table containing the numbers of accidents and the frequencies of each number looks like:

Accidents	Count
0	296
1	74
2	26
3	8
4	4
5	4
6	1
8	1
N=	414

(b) The estimated proportion of machinists who suffered at least one injury is 0.285. An approximate 99% confidence interval for this proportion is $(0.228, 0.342)$.

Solution to Activity 19

(a) The difference between the proportions who never snore in the two groups is

$$\frac{1355}{2374} - \frac{24}{110} \simeq 0.353.$$

Thus, in this sample, the proportion who never snore is much greater in people without heart disease than in people with heart disease. (You were asked to estimate the difference this way round, but had you taken the difference between the proportion of people with heart disease who never snore and the proportion without heart disease who never snore, you'd have obtained -0.353.)

(b) An approximate 95% confidence interval for the difference between the proportion of people without heart disease who never snore and the proportion with heart disease who never snore is $(0.273, 0.432)$. (The confidence interval for the difference the other way round would have been $(-0.432, -0.273)$.)

Let p_1 be the proportion of people without heart disease who never snore and p_2 the proportion of people with heart disease who never snore. The 95% confidence interval for $p_1 - p_2$ includes only positive values. (Similarly, the 95% confidence interval for $p_2 - p_1$ includes only negative values.) This suggests that p_1, the proportion of people without heart disease who never snore, is higher than p_2, the proportion of people with heart disease who never snore. Note that you cannot infer from this that heart disease causes snoring; it is just that people with heart disease seem to be more likely to snore, but this could be for any number of reasons.

Solution to Activity 20

(a) The estimated difference between the proportions is -0.273, with 95% confidence interval $(-0.388, -0.158)$.

(b) The 95% confidence interval includes only negative values. This suggests that p_1, the proportion of sewerage workers who are infected and don't have children, is lower than p_2, the proportion of sewerage workers who are infected and do have children. This in turn suggests that sewerage workers with children are more likely to have been infected with hepatitis A than sewerage workers without children. Note that you cannot infer from this that the workers are getting hepatitis A from their children! There may be another explanation for this difference. For instance, having more children is likely to be related to age, and hence to the duration of exposure to environmental risks.

Solution to Activity 21

(a) A normal probability plot is shown in Figure 13.

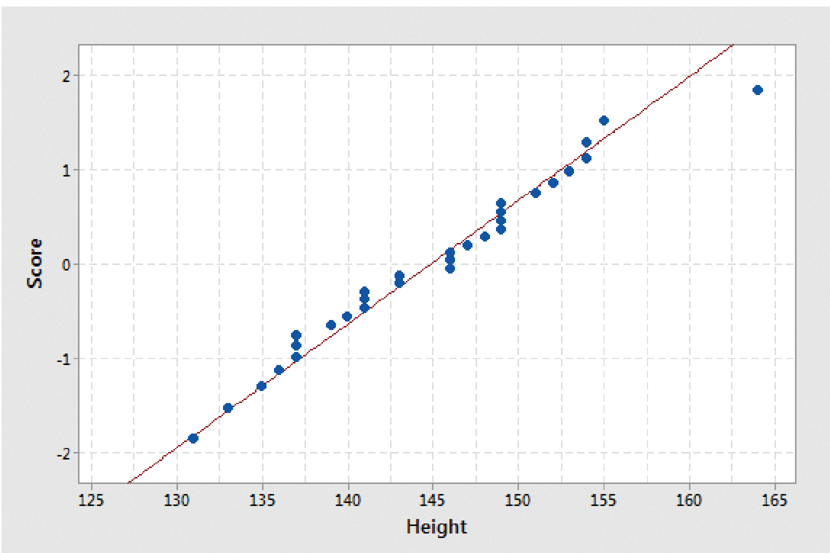


Figure 13 A normal probability plot of heights

The points lie reasonably close to a straight line, so the normality assumption appears reasonable.

(b) The estimated mean height is 144.87 cm. A 90% t-interval for the mean height in cm is $(142.50, 147.23)$.

Solution to Activity 22

(a) The sample means and standard deviations may be obtained using **Stat > Basic Statistics > Display Descriptive Statistics. . .** . The difference between the means (Etruscans − Italians) is $143.77 - 132.44 = 11.33$ mm. The sample standard deviation for the Etruscan skulls is 5.97 and for the Italian skulls is 5.75. The ratio of the larger sample variance to the smaller is $(5.97/5.75)^2 \simeq 1.08$. Since this is less than 3, you can assume that the underlying variances are equal.

Graph > Probability Plot. . .

(b) Figure 14 shows the normal probability plots for `Etruscans` and `Italians`.

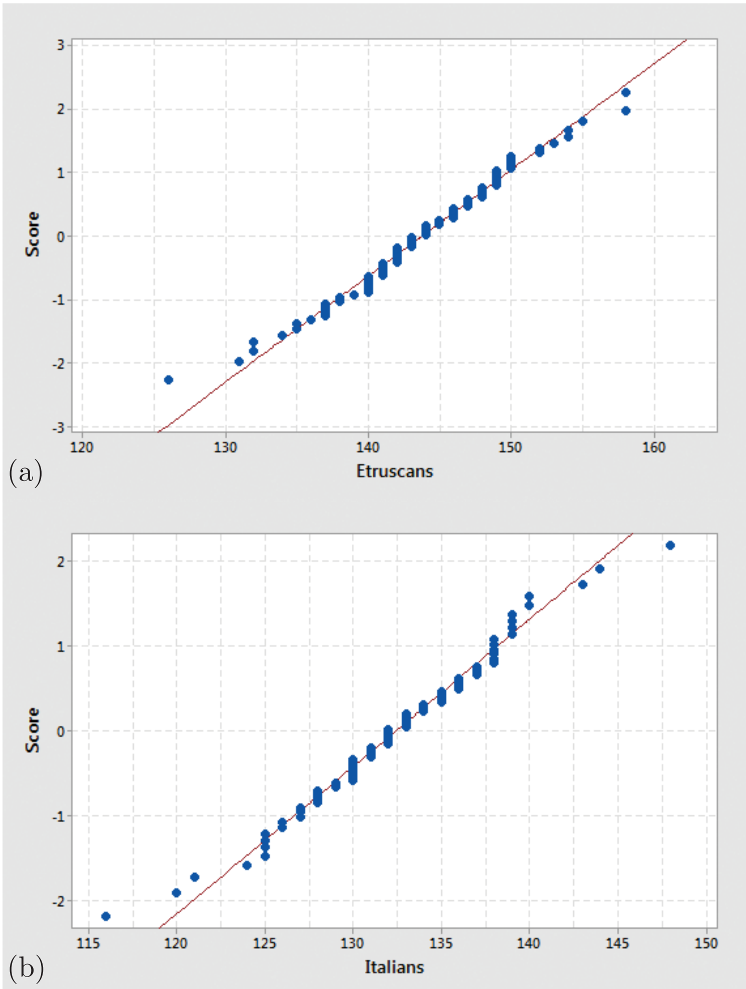


Figure 14 Normal probability plots for (a) Etruscans and (b) Italians

The plots suggest that the normality assumption is reasonable for both populations.

(c) A 95% two-sample t-interval for the difference in mm between the mean skull breadth of Etruscans and that of modern Italians, calculated assuming equal variances, is $(9.454, 13.208)$.

(d) When it is not assumed that the variances are equal, the confidence interval is $(9.460, 13.202)$.

Solution to Activity 25

(a) To create the variable `Difference`, in the **Calculator** dialogue box enter `Difference` in the **Store result in variable** field, and `Cross - Self` in the **Expression** field.

In the **One-Sample t for the Mean** dialogue box, select **One or more samples, each in a column** from the drop-down list at the top, enter `Difference` underneath, select **Perform hypothesis test** and enter `0` in the **Hypothesized mean** field. Here we have a two-sided test, so in the **One-Sample t: Options** dialogue box, select `Mean` $\neq$ `hypothesized mean` from the **Alternative hypothesis** drop-down list.

The p-value given by Minitab is 0.050. This (just) provides moderate evidence against the null hypothesis. There is thus moderate evidence that the plants differ in mean height. Indeed, since the sample mean difference is 20.93, the data suggest that cross-fertilised plants are taller on average than self-fertilised plants.

(b) Minitab performs exactly the same test as in part (a) and so we come to exactly the same conclusion.

Solution to Activity 27

Stat > Basic Statistics > 1 Proportion...

(a) Follow the instructions for obtaining a p-value given in Activity 26, entering `111` in the **Number of events** field (since 111 of the respondents in Northern Ireland were living with their parents) and `307` in the **Number of trials** field (because there were 307 people in this sample).

The p-value given by Minitab is 0.000. This means that $p < 0.01$. So there is strong evidence against H_0 that $p_{NI} = 0.25$. The data suggest that in Northern Ireland, the proportion of young adults aged between 20 and 34 living with their parents is not 0.25. Further, since the test statistic is positive, or because the sample proportion is approximately 0.36, the data suggest that the proportion living with their parents in Northern Ireland is higher than 0.25.

(b) Follow the instructions for obtaining a p-value given in Activity 26, entering `86` in the **Number of events** field (since 86 of the respondents in Scotland were living with their parents) and `344` in the **Number of trials** field (because there were 344 people in this sample).

The p-value given by Minitab is 1.000. At first sight this might seem a strange value! However, the observed value of the test statistic is 0.00 (since $\widehat{p}$, given by Minitab under `Sample p`, equals 0.25), which is exactly in the centre of the $N(0,1)$ distribution. Therefore *all* values

of the test statistic z will be 'at least as extreme as' the observed value of the test statistic because

$$p = P(Z \leq 0) + P(Z \geq 0) = 1.$$

So for Scotland there is no evidence that the proportion of young adults aged between 20 and 34 years old is not 0.25.

Solution to Activity 29

The theoretical percentage is 67.0%.

Since random data are generated by the animation, it is not possible to predict the precise results that you will obtain. However, you probably found that somewhere between 65% and 70% of your samples gave values of T in the rejection region. Putting this another way, the alternative hypothesis is true (because $\mu \neq 0$), and for between 65% and 70% of samples, the test detects this correctly by rejecting the null hypothesis. In the remaining 30% to 35% of samples, the observed value of the test statistic does not fall in the rejection region and the null hypothesis is not rejected, so a Type II error has been made.

Solution to Activity 30

The theoretical percentage is 99.8%.

In this case, the difference between the hypothesised mean of 0 and the actual population mean, now 1, is larger than it was in Activity 29, while all other settings remain the same. So you might have thought that the test is more likely to distinguish between the true value of the population mean (1) and the hypothesised value (0), and that more of the observed values of the test statistic will fall in the rejection region. You would be right. You probably found that, for over 99% of the samples, the observed value of the test statistic fell in the rejection region. So, when the true mean is further away from the hypothesised mean, the probability of making a Type II error decreases.

Solution to Activity 32

The output in the Session window gives the required sample size as 215 for the power required (`Target Power`). Minitab also reports that the power of the test for a sample size of 215 (`Actual Power`) is, in fact, $0.901079 \simeq 0.901$, which is slightly higher than the power that you asked for. This is the power for sample size 215, and the fact that the actual power is slightly larger than the target power means that the sample size was rounded up to the nearest whole number.

Solution to Activity 33

(a) When the sample size is 25, the power calculated using the z-test was $0.259511 \simeq 0.260$. When using a t-test the power is $0.250485 \simeq 0.250$. So these two values are similar.

(b) When the sample size is 10, the power calculated using the z-test is $0.155674 \simeq 0.156$. When using a t-test the power is $0.144818 \simeq 0.145$. So this time the two values are further apart.

Solution to Activity 34

(a) The sample size calculated for the z-test was 215. The sample size calculated for the t-test is 216. The sample sizes calculated are very similar.

(b) The sample size is calculated to be 9 for the z-test when $d = 10$. The sample size is calculated to be 11 for the t-test when $d = 10$. The sample sizes calculated are less similar, especially in relative terms. This is because we are dealing with small sample sizes where the differences between the normal and $t(n-1)$ distributions are more pronounced.

Solution to Activity 35

Use **Calc > Calculator...** to create a column called `Difference` containing the differences, `Glaucomatous - Normal`.

The output for the Wilcoxon signed rank test begins by stating the hypotheses for the test. Then you are told that there were 8 values in the sample (`N`), but that only 7 of them were used for the test (`N for test`). (This is because, as in Unit 10, sample differences of 0 are excluded from the calculations. Here there is one zero value, so Minitab excludes it and proceeds with the test as if only the other 7 sample values had been observed.)

The observed value w_+ of the Wilcoxon signed rank test statistic is 7.5, as calculated in Unit 10. The p-value is 0.310. This is not the same as the exact p-value of 0.344 given in Unit 10; the reason for the difference is explained in the text following this activity. None the less, such a high p-value provides little or no evidence against the null hypothesis.

See Example 3 of Unit 10.

Finally, an estimate of the population median is given. Calculation of this estimate is not really part of the hypothesis testing procedure. However, it happens to be the case that the value of the estimate, -4, is not quite the same as the sample median (of all the differences, zero or not), which is -3. The reason for this discrepancy is also explained briefly in the text following this activity.

It is not the sample median of the seven non-zero differences either, which is -6.

Solution to Activity 36

Following the same instructions for carrying out the Wilcoxon signed rank test as given for Activity 35, the p-value given by Minitab is 0.721. This is close to the exact p-value of 0.709 quoted in Unit 10.

The conclusion is the same as in Unit 10: the Wilcoxon signed rank test provides little or no evidence against the hypothesis that the population median difference between percentages of time a foetus spends moving pre- and post-test are zero.

Solution to Activity 37

(a) The appropriate hypotheses for the test are either:

$$H_0 : m = 0, \quad H_1 : m > 0,$$

where m is the population median difference between lesion counts

(A − B), or

$$H_0 : m = 0, \quad H_1 : m < 0,$$

where m is the population median difference between lesion counts (B − A). In each case, the direction of the inequality in H_1 is due to the expectation that preparation A would lead to more lesions than preparation B.

Graph > Probability Plot...

(b) Use **Calc > Calculator...** to create a column called `Difference` containing the differences, `Alesions` − `Blesions`. A normal probability plot of `Difference` is shown in Figure 15.

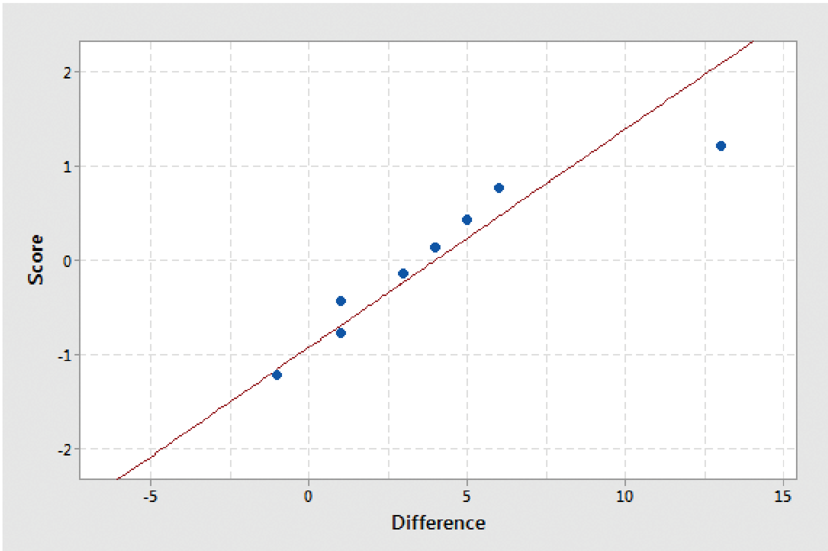


Figure 15 A normal probability plot for the differences

Most of the points lie quite close to a straight line, but the largest value (a difference of 13 for the first leaf) could be argued to be an outlier. Even though the dataset is rather small, this does cast some doubt on whether the underlying distribution is normal. (Another possibility is that the rest of the distribution is normal, but that this particular leaf was unusual in some way and/or had its lesions counted or recorded incorrectly.)

Stat > Nonparametrics > 1-Sample Wilcoxon...

(c) The test is performed by obtaining the **1-Sample Wilcoxon** dialogue box, entering `Difference` into the **Variables** field, choosing **Test median** and entering `0` in the **Test median** field, and setting the **Alternative** drop-down list to `greater than`.

The test statistic for the Wilcoxon signed rank test is 34 and the p-value is 0.015. Since $0.01 < p < 0.05$, this provides moderate evidence against the null hypothesis and in favour of the alternative hypothesis that the median difference between lesion counts is greater than 0; that is, there is moderate evidence that preparation A tends to produce more lesions on a leaf than does preparation B.

However, there are some grounds for caution about the use of the Wilcoxon signed rank test in this context. First, the possible outlier

identified in part (b) casts some doubt on the assumption of symmetry of the underlying distribution. Second, the sample size is only 8, and there is a set of three tied values in the ranks. Thus the basic normal approximation used in Unit 10 to calculate the p-value may not be very accurate; the approximate p-value calculated by Minitab includes a refinement, but it is possible that this approximation may not be very accurate either.

In fact, the exact p-value, found using other software, is 0.0156 for these data. So, in fact, the approximation is good in this case.

Solution to Activity 38

(a) For this Mann–Whitney test, the output begins with the sample sizes and sample medians for the two samples. Then it provides a 'point estimate' and a confidence interval for the difference between the population medians. (You should ignore these, since our primary interest here is in the hypothesis test.)

Next, the value of a quantity called `W` is given. You might have recognised this as the Mann–Whitney test statistic, which is called U_A in Unit 10. Here it takes the value 185, the same as in the unit.

Then Minitab says which test has been performed. It states the hypotheses in terms of the group population medians, η_1 and η_2; in Unit 10, we state the hypotheses in terms of the population difference in location ℓ. The two are the same if we equate ℓ with $\eta_1 - \eta_2$. Then, the null hypothesis can be written equivalently as

$$H_0 : \eta_1 = \eta_2$$

here or as

$$H_0 : \ell = 0$$

in Unit 10. Similarly, the (two-sided) alternative hypothesis can be written equivalently as

$$H_1 : \eta_1 \neq \eta_2$$

(Minitab) or

$$H_1 : \ell \neq 0$$

(Unit 10).

In the same row of output, Minitab tells us that the test 'is significant at 0.0025'; this is its rather unusual way of saying that the p-value for the test is 0.0025. Note that in Example 7 of Unit 10, it was stated that the exact p-value is 0.0015 and the p-value using the normal approximation was calculated to be 0.0022: the reason for the difference in p-values is explained in the text following this activity.

Finally, a p-value after adjusting for ties is given. This is something provided by Minitab, the details of which need not concern you. In this case, there are no ties in the data, so this 'adjusted' p-value is the same as the unadjusted p-value anyway.

The p-value is very small, so there is strong evidence that the two populations differ in location, that is, that on average patients classified as psychotic have different dopamine activities to patients

classified as non-psychotic. Indeed, it appears that the psychotic patients tend to have higher dopamine activities (since the sample median is higher for the psychotic patients). This is the same conclusion as was arrived at in Unit 10.

(b) When the samples are entered the other way round, the sample sizes and sample medians are the same (but given in the other order).

The point estimate and confidence interval for the difference between the population medians are the same as before, except that the signs have changed. This is because Minitab is now labelling the two populations in the opposite order and calculating the differences the other way round.

The test statistic is now given as 140 instead of 185; this is again because the samples are labelled the other way round. Minitab always reports, as the test statistic, the rank sum for the **First Sample**. In Unit 10, you saw that the rank sums for the two samples were 185 for those judged psychotic and 140 for those judged non-psychotic. So, in this analysis, the reported test statistic is the rank sum for those judged non-psychotic, because they were entered as the **First Sample**.

Since this is a two-sided test, the difference in order of the two samples should make no difference to the amount of evidence against the null hypothesis, and hence the p-value should not depend on the order in which the samples were entered. Indeed, the p-value for this test is the same as that for the first Mann–Whitney test.

Solution to Activity 39

The appropriate Mann–Whitney test can be set up in Minitab in two different ways. In the **Mann-Whitney** dialogue box, if you enter `Concinna` in the **First Sample** field and `Picipes` in the **Second Sample** field, then the value of the test statistic is reported to be 127.0. If you enter `Picipes` in the **First Sample** field and `Concinna` in the **Second Sample** field, then the value of the test statistic is reported to be 83.0. This is because the value reported is always the rank sum for the data entered as **First Sample**. In either case, the p-value is given by Minitab as 0.1041 or, when adjusted for ties, the very similar 0.1031.

Since $p > 0.1$, the conclusion is that there is little or no evidence against the null hypothesis that the difference in locations of the population distributions of the widths of the first joints of the second tarsus is zero (although you should note that p is only very slightly larger than 0.1, and so you might emphasise 'little' over 'no'). So, there is little or no evidence that we can discriminate between these two species of beetle on the basis of this joint width measurement.

Solution to Activity 40

The null hypothesis is that the population distributions of OSA scores in the two types of society are the same (in terms of location and everything else). Here, you are asked to perform a one-sided test, investigating the

hypothesis that OSA scores tend to be higher in societies where oral explanations of illness are present than in societies where oral explanations of illness are absent.

The appropriate Mann–Whitney test can be set up in Minitab in two different ways using the **Mann-Whitney** dialogue box.

Stat > Nonparametrics > Mann-Whitney...

If you enter `Present` in the **First Sample** field and `Absent` in the **Second Sample** field, then you must set the **Alternative** drop-down list to `greater than` (meaning that 'present' scores tend to be greater than 'absent' scores).

On the other hand, if you enter `Absent` in the **First Sample** field and `Present` in the **Second Sample** field, then you need to set the **Alternative** drop-down list to `less than` (meaning that 'absent' scores tend to be smaller than 'present' scores).

A different value of the test statistic is reported in these two cases (580 for the first, 200 for the second). This is again because the value reported is always the rank sum for the data entered as **First Sample**. However, in calculating the corresponding p-value, Minitab uses a different tail of the null distribution in the two cases, which leads to the same p-value whichever way round we work.

If you are interested, the details of how this works out in the current case are given in this paragraph. (If not, please feel free to skip to the next paragraph.) In the first case, corresponding to the 'greater than' alternative, the p-value is

$$p = P(U \geq 580)$$

where U is the random variable representing the sum of ranks for the first sample, which in this case is 'present'. In the second case, corresponding to the 'less than' alternative, the p-value is

We avoid calling U and V U_A and U_B because the labels A and B swap around between cases.

$$p = P(V \leq 200)$$

where V is the random variable representing what is again the sum of ranks for the first sample, but this is now the 'absent' sample. However, recall that if n_A and n_B denote the sample sizes of the two samples, then the sum of the ranks of the two samples must equal

$$\tfrac{1}{2}(n_A + n_B)(n_A + n_B + 1) = \tfrac{1}{2} \times 39 \times 40 = 780.$$

That is, $U + V = 780$ or $U = 780 - V$ which means that

$$p = P(U \geq 580) = P(780 - V \geq 580) = P(780 - 580 \geq V) = P(V \leq 200).$$

Therefore, the p-value is the same whichever set-up is used.

The p-value given by Minitab is 0.0003. (In this case, although there are many ties, the adjustment for ties makes no difference at all to the p-value.)

The conclusion is that there is strong evidence against the null hypothesis. There is strong evidence that societies where oral explanations of illness are present tend to have higher oral socialisation anxiety scores than societies where oral explanations of illness are absent.

Solutions to exercises

Solution to Exercise 1

Calc > Probability Distributions > Normal...

In the **Normal Distribution** dialogue box, enter `100` for the mean of the normal distribution and `15` for the standard deviation. Use **Cumulative probability** to answer parts (a) and (b), and **Inverse cumulative probability** to answer parts (c) and (d).

(a) If X is a random variable representing the score of a randomly selected person, then the proportion of the population recording scores of 120 or more is given by

To find $F(120)$, enter `120` into the **Input constant** field.

$$P(X \geq 120) = 1 - F(120) = 1 - 0.908789 \simeq 0.0912.$$

(b) The proportion of the population recording scores between 110 and 130 is given by

$$P(110 \leq X \leq 130) = F(130) - F(110) = 0.977250 - 0.747507 \simeq 0.2297.$$

(c) The score required is the 0.15-quantile of $N(100, 15^2)$:

To find $q_{0.15}$, enter `0.15` into the **Input constant** field.

$$q_{0.15} = 84.4535 \simeq 84.5.$$

(d) The score required is the 0.95-quantile of $N(100, 15^2)$:

$$q_{0.95} = 124.673 \simeq 124.7.$$

Solution to Exercise 2

Graph > Probability Plot...

The normal probability plots for `Coin2`, `Coin3` and `Coin4` are shown in Figure 16.

The data lie roughly along the straight lines for each graph, although there is a suggestion of a systematic pattern around the straight line for `Coin2` and a suggestion of a curve for `Coin3`. So, a normal distribution is a plausible model for `Coin4`, but is perhaps questionable for `Coin2` and `Coin3`. However, the sample sizes are small – especially for `Coin3` – and so these plots need to be treated with caution. It is, for example, never possible to make firm conclusions on the basis of just four data points.

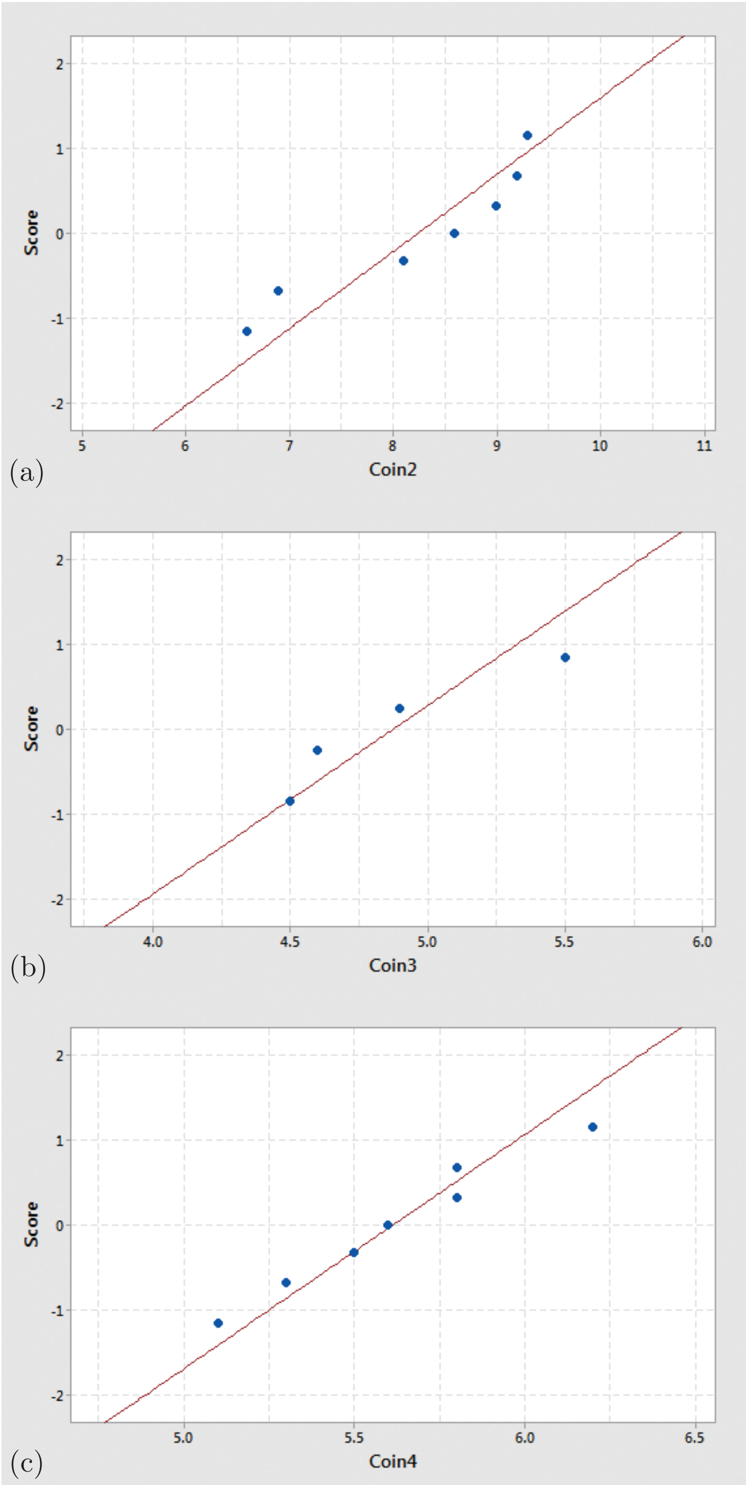


Figure 16 Normal probability plots for (a) `Coin2`, (b) `Coin3` and (c) `Coin4`

Solution to Exercise 3

(a) Use **Stat > Basic Statistics > 1 Proportion...**, selecting **Summarized data** from the drop-down list at the top. Remember that in the **One-Sample Proportion: Options** dialogue box, you need to select **Normal approximation** from the **Method** drop-down list. The approximate 95% confidence intervals are $(0.464, 0.736)$ for sewerage workers with frequent exposure to raw sewage, and $(0.210, 0.341)$ for workers with infrequent or no exposure to raw sewage.

(b) Use **Stat > Basic Statistics > 2 Proportions...**, again selecting **Summarized data** from the drop-down list at the top. In the **Two-Sample Proportion: Options** dialogue box, make sure that **Estimate the proportions separately** is selected from the **Test method** drop-down list. The estimated difference between the proportions (frequently exposed minus the others) is 0.325, with approximate 95% confidence interval $(0.174, 0.476)$.

(c) If the entire study were repeated independently a large number of times, and on each occasion a 95% confidence interval for the difference in proportions were calculated, then about 95% of these intervals would contain the population value of the difference between proportions of sewerage workers with frequent exposure to raw sewage and sewerage workers with infrequent or no exposure to raw sewage having had hepatitis A. The interval includes only positive values, suggesting that frequently exposed workers are at higher risk of hepatitis A infection than unexposed or infrequently exposed workers.

Solution to Exercise 4

(a) The variable `Ratio` is constructed using **Calc > Calculator...**. Enter `Weight/(Height - 100)` in the **Expression** field.

Graph > Probability Plot...

(b) A normal probability plot for the ratios is shown in Figure 17.

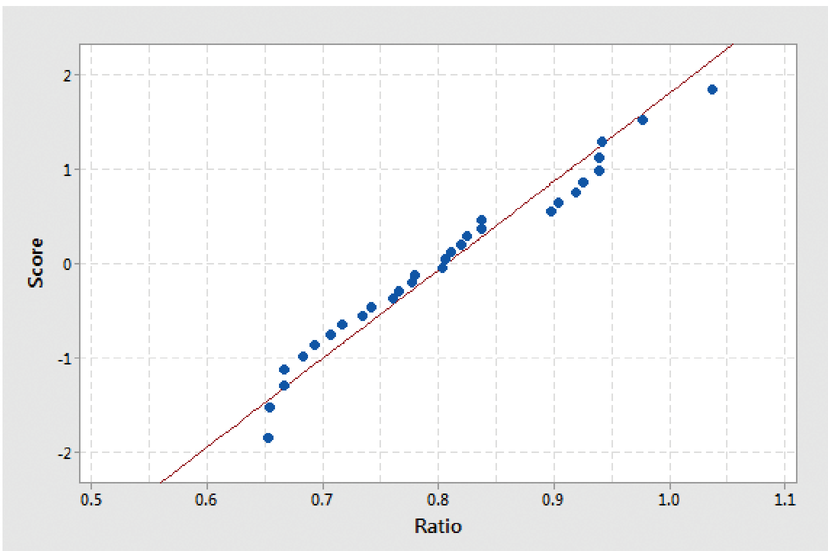


Figure 17 A normal probability plot for the ratios

The points lie roughly along a straight line, although the pattern is not perfect, neither in the lower tail nor in the gap from `Ratio` values of about 0.85 to 0.9. You can argue either way about the reasonableness or otherwise of the normality assumption on the basis of this plot!

(c) Use **Stat** > **Basic Statistics** > **1-Sample t...**, selecting **One or more samples, each in a column** from the drop-down list at the top and entering `Ratio` in the field underneath. The 95% t-interval for the population mean of the ratios is $(0.768, 0.847)$. The sample mean and sample standard deviation of `Ratio` are 0.8076 and 0.1065, respectively.

(d) Use **Stat** > **Basic Statistics** > **1-Sample Z...**, again selecting **One or more samples, each in a column** from the drop-down list at the top and entering `Ratio` in the field underneath. Enter the standard deviation `0.1065` (as calculated in part (c)) in the **Known standard deviation** field. The approximate 95% confidence interval thereby obtained is $(0.770, 0.846)$. The approximate confidence interval is a little narrower than the exact confidence interval (which assumes normality), but the two are the same to two decimal places. It seems that the reasonableness or otherwise of the normality assumption is not of crucial importance in terms of these confidence intervals.

Solution to Exercise 5

(a) The sample standard deviation for males is 7.520 and for females is 7.365.

Stat > **Basic Statistics** > **Display Descriptive Statistics...**

(b) Using **Stat** > **Basic Statistics** > **1-Sample Z...**, select **One or more samples, each in a column** from the drop-down list at the top and enter `Male` in the field underneath. Enter the value of the sample standard deviation for `Male`, 7.52, in the **Known standard deviation** field, select **Perform hypothesis test** and enter `50.7` in the **Hypothesized mean** field.

Minitab gives the value $p = 0.000$, so $p < 0.01$ and there is strong evidence against the null hypothesis that the mean pass rate for the driving practical test nationally for males over the period April 2014–March 2015 is the same as the national pass rate for males over the same period in the previous year. Further, since the observed test statistic, z, is positive, it suggests that the mean pass rate for males nationally over the period April 2014–March 2015 is higher than the national pass rate for males over the same period in the previous year.

(c) Using **Stat** > **Basic Statistics** > **1-Sample Z...**, select **One or more samples, each in a column** from the drop-down list at the top and enter `Female` in the field underneath. Enter the value of the sample standard deviation for `Female`, 7.365, in the **Known standard deviation** field, select **Perform hypothesis test** and enter `43.8` in the **Hypothesized mean** field.

Minitab again gives the value $p = 0.000$ so that $p < 0.01$. There is therefore also strong evidence against the null hypothesis that the mean pass rate for the driving practical test nationally for females

over the period April 2014–March 2015 is the same as the national pass rate for females over the same period in the previous year. Further, since the observed test statistic, z, is positive, it suggests that the mean pass rate for females nationally over the period April 2014–March 2015 is higher than the national pass rate for females over the same period in the previous year.

(d) Taken together, the results suggest that the mean pass rate for the driving practical test nationally during the period April 2014–March 2015 was higher than the national pass rate during the same period in the previous year, and this higher pass rate was evident for both males and females.

Solution to Exercise 6

Stat > Basic Statistics > Display Descriptive Statistics...

(a) The sample mean is 78.95 inches and the sample standard deviation is 5.69 inches.

(b) Using **Stat > Basic Statistics > 1-Sample t...**, select **One or more samples, each in a column** from the drop-down list at the top and enter `Height` in the field below. Select **Perform hypothesis test** and enter `82` in the **Hypothesized mean** field. In the **One-Sample t: Options** dialogue box, make sure the **Confidence level** field contains `95` and that `Mean ≠ hypothesized mean` is selected from the **Alternative hypothesis** drop-down list.

The 95% confidence interval for μ is $(76.20, 81.69)$ inches, and the p-value is 0.031.

(c) The p-value is such that $0.01 < p \leq 0.05$ and so there is moderate evidence against the null hypothesis that the mean height of basketball players is 82 inches. Since the sample mean is 78.95 inches, the evidence suggests that the mean height of basketball players is less than 82 inches.

(d) Because $p = 0.031 < 0.05$, a hypothesis test at the 5% significance level would have been rejected. This is in agreement with the fact that the hypothesised value of μ, 82 inches, lies outside the 95% confidence interval of $(76.20, 81.69)$ inches. And the values in the confidence interval agree with the suggestion in the solution to part (c) that the mean height of basketball players is less than 82 inches.

Solution to Exercise 7

(a) Using **Stat > Basic Statistics > 1-Sample t...**, select **One or more samples, each in a column** from the drop-down list at the top and enter `Ltreatment` in the field below. Select **Perform hypothesis test** and enter `0` in the **Hypothesized mean** field. In the **One-Sample t: Options** dialogue box, select `Mean > hypothesized mean` from the **Alternative hypothesis** drop-down list.

Minitab gives the test statistic as 3.68, with a (one-sided) p-value of 0.003. Since $p < 0.01$, there is strong evidence against the null

hypothesis and in favour of the alternative hypothesis that the underlying mean sleep gain of patients who have taken this drug is greater than zero.

(b) A t-test for this treatment is carried out as in part (a), only using the variable `Dtreatment`. For this treatment, Minitab gives the test statistic as 1.33, with a (one-sided) p-value of 0.109. Since $p > 0.1$, there is little or no evidence against the null hypothesis that the mean sleep gain is zero when using this drug.

(c) Using **Stat > Nonparametrics > 1-Sample Wilcoxon...**, enter both variables `Ltreatment` and `Dtreatment` into the **Variables** field, enter `0` in the **Test median** field, and select `greater than` from the **Alternative** drop-down list.

For L-hyoscyamine hydrobromide (`Ltreatment`), Minitab gives a (one-sided) p-value of 0.005. Since there are ten sample values and only one pair of tied ranks, the p-value for the Wilcoxon signed rank test is likely to be a good approximation. (In fact, the exact p-value is 0.003.) There is strong evidence against the null hypothesis that this drug produces no sleep gain on average. That is, there is strong evidence that this drug increases sleep time.

For D-hyoscyamine hydrobromide (`Dtreatment`), Minitab gives a (one-sided) p-value of 0.172. There is little or no evidence against the null hypothesis that this drug produces no sleep gain on average.

(d) Use **Calc > Calculator...** to create a variable `Difference` entering `Ltreatment - Dtreatment` in the **Expression** field. A probability plot for `Difference` is shown in Figure 18.

Graph > Probability Plot...

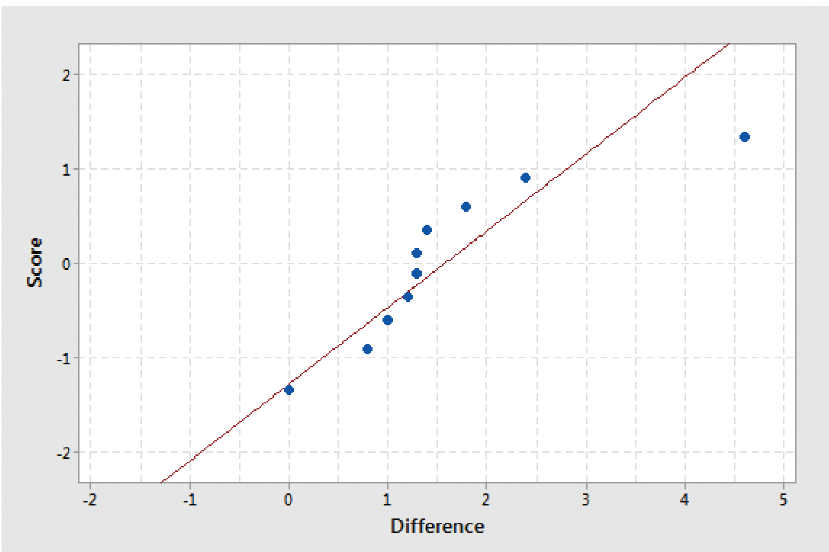


Figure 18 A normal probability plot for the differences

The pattern of points looks very curved, throwing doubt on the normality of the distribution. There also seems to be a high outlier.

(e) The appropriate test is again the Wilcoxon signed rank test, but this time in two-sided form using the differences rather than the individual samples.

Using **Stat > Nonparametrics > 1-Sample Wilcoxon...**, enter the variable `Difference` in the **Variables** field, enter `0` in the **Test median** field, and select `not equal` from the **Alternative** drop-down list.

The (two-sided) p-value for the test is reported as 0.009 by Minitab. Since $p < 0.01$, there is strong evidence against the null hypothesis that the two drugs are equally effective. Since the sample median difference in sleep gain (L-hyoscyamine minus D-hyoscyamine) is greater than zero, the data suggest that L-hyoscyamine hydrobromide is more effective at prolonging sleep.

The sample size is not large and the outlier in Figure 18 gives some doubt about the symmetry of the underlying distribution, so this p-value may not be entirely accurate. Nevertheless, a p-value as small as this still suggests that there is evidence against the null hypothesis.

Solution to Exercise 8

The null hypothesis is that the locations of the distributions of median shot lengths are the same in the talkies of Laurel and Hardy as in their silent films. We are not given any prior clue as to which set of shots might be longer on average, and so use a two-sided test.

Stat > Nonparametrics > Mann-Whitney...

The appropriate test is the Mann–Whitney test. It can be set up in Minitab in two different ways using the **Mann-Whitney** dialogue box, depending on which sample you enter in the **First Sample** field. If it's `Silent`, then the value of the test statistic is reported to be 134.0; if it's `Sound`, the value of the test statistic is reported to be 394.0. Either way, the p-value for the two-sided test is given by Minitab as 0.0134 or, when adjusted for ties, the very similar 0.0128.

Since $0.01 < p \leq 0.05$, the conclusion is that there is moderate evidence against the null hypothesis that the median shot lengths are the same in the two types of film. Looking at the data, it seems that there is moderate evidence that median shot lengths tend to be longer in Laurel and Hardy's talkies than in their silent films.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 6: Taken from Old Maps

Page 7: Jerod Harris / Stringer / Gettyimages

Page 10: Taken from Wikimedia.org

Page 11: Catalog No 50/2347 Field No 47. Courtesy of the Division of Anthropology. American Museum of Natural History.

Page 12: Courtesy of Jiaya Jia

Page 23 top: Mark Parisi/Off The Mark Cartoons

Page 23 bottom: ‘Women war machinists engaged in making artillery shells at Woolwich Arsenal,’ by Unknown. The Imperial War Museum via First World War Poetry Digital Archive, accessed May 3, 2017, http://ww1lit.nsms.ox.ac.uk/ww1lit/collections/item/3716.

Page 25: Fanch / iStock / Getty Images Plus

Page 28: Norman Einstein This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 29: Brian Jackson / 123rf.com

Page 31: ValentynVolkov / iStockphoto / Getty Images

Page 32: FreeBigPictures.com

Page 40: www.ranker.com This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 45: Courtesy of Grazia Trebbi

Page 46: Taken from designertravel.co.uk

Page 47: Lech Borowiec

Page 48: Image © 2017, Peter Henderson, Pisces Conservation Ltd

Page 51: Peepo / iStock / Getty Images Plus

Page 52: General Photographic Agency / Hulton Archive / Getty Images / Universal Images Group

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index